AD-A049 826 AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER LACK--ETC F/G 5/9
A METHODOLOGICAL STUDY OF FORCED-CHOICE PERFORMANCE RATING.(U)
MAY 51 R W HIGHLAND, J R BERKSHIRE

UNCLASSIFIED RB-51-9 NL

1 of 1
ADA049 826

END
DATE FILMED
3-78
DDC

AIR TRAINING COMMAND

# HUMAN RESOURCES RESEARCH CENTER















9 DEC 1956

(9) *Research Bulletin 51-9*

(12) 57p.

(6) ## A METHODOLOGICAL STUDY OF FORCED-CHOICE PERFORMANCE RATING

*by*

(10) RICHARD W. HIGHLAND
and
JAMES R. BERKSHIRE













Lackland Air Force Base
San Antonio, Texas

(11) May 1951

# A METHODOLOGICAL STUDY OF FORCED-CHOICE PERFORMANCE RATING

PROJECT NO. 21-07-009



By

RICHARD W. HIGHLAND

and

JAMES R. BERKSHIRE



TECHNICAL TRAINING RESEARCH LABORATORY

DETACHMENT NO. 3

HUMAN RESOURCES RESEARCH CENTER

CHANUTE AIR FORCE BASE, ILLINOIS



RESEARCH BULLETIN 51-9

May 1951

SUBMITTED BY:

THOMAS W. HARRELL
*Director of Research*
Technical Training Research Laboratory

ACKNOWLEDGMENTS

Captain Paul T. Dupell, Lt. Howard J. Zeman, and S/Sgt. Daniel E. Lacey assisted in obtaining the rankings and ratings of instructors. T/Sgt. John E. Burks, Jr. and S/Sgt. Lacey computed the preference, favorableness, and discrimination indices.

As Chief of the Technical Services Division, Mr. George B. Simon reviewed the over-all plans and provided interpretations of the analysis of data. In addition, Mr. Simon and Mr. George G. Burgess made valuable contributions in the general planning of the project and particularly in the analysis of the data. The development and supervision of the detailed analysis were accomplished at various stages by Mr. Donald B. Devoe, Mr. Burgess, and Mr. Risdon J. Westen.

The figures were prepared by Cpl. Richard A. Spoor and Cpl. Carl J. Nordquist.

The manuscript was reviewed by Dr. Lee J. Cronbach and associates in the Bureau of Research and Service, University of Illinois.

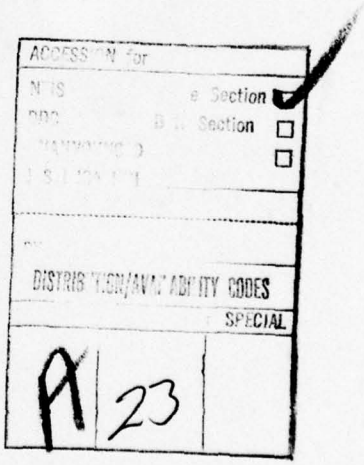

TABLE OF CONTENTS

## LIST OF TABLES



## LIST OF FIGURES

## ABSTRACT

While there are general similarities among the different kinds of forced-choice rating forms reported in the psychological literature, there are also basic differences among them. When the Human Resources Research Center of the Air Training Command was requested to develop a new rating form for Air Force technical school instructors, an opportunity for a comparative study of these basic differences was offered.

The objective of the present study was a relative evaluation of six different kinds of forced-choice performance rating forms with respect to validity, reliability, biasability, and the degree to which raters liked different forms. The kinds of forced-choice forms used in the investigation varied as to the content of the blocks of statements and as to the directions which were given the raters.

About 2300 ratings were obtained from the six Air Force bases associated with technical training. Item-analysis keys were developed using three bases as one sample and the other three bases as another sample for purposes of cross-validation. Validities were obtained by correlating scores on the various forms with proficiency rankings as assigned by instructor supervisors. Using the average validities for five different scoring keys which were developed, validities ranging from .53 to .69 were obtained under conditions of cross-validation. Reliabilities were obtained by using the split-half technique, attempting to equalize the two halves of the form with respect to block validities. The stepped-up reliability coefficients ranged from .68 to .96. When these were further adjusted to make all forms equivalent in length to the longest form, the resulting reliability coefficients ranged from .91 to .97.

Of the six forms used in this project, a form made up of blocks containing four favorable-appearing statements from which the rater was asked to choose the two which were most descriptive of the ratee gave generally superior results. This form had highest average validity (.68) and satisfactory reliability (.93), was least susceptible to deliberate attempts to give high scores, and was one of the two forms best liked by the raters.

Forms made up of blocks of four or five statements, of which some appear favorable and some unfavorable, and from which the rater is asked to choose the most descriptive statement and the least descriptive statement, have been widely used in military and industrial situations. In the present experiment two forms constructed in this manner showed a relatively strong tendency to produce negatively skewed distributions when the raters were told to give as high a score as possible. The

validities obtained for these two forms were the lowest (.53 and .56) of those forms tested, although the forms yielded the highest reliabilities (.96 and .97).

## Conclusions

Of the forms used in this experiment, those in which the blocks were composed of four favorable-appearing statements, from which the rater was to choose either the two most descriptive, or the most and the least descriptive, were generally superior.

The inclusion of both favorable and unfavorable statements in the same block appears to be an inferior method of constructing forced-choice forms.

# A METHODOLOGICAL STUDY OF FORCED-CHOICE PERFORMANCE RATING

## INTRODUCTION

It is a part of the American culture that individuals should be rewarded for effectiveness in doing a job. It is also a part of our concept of efficiency that individuals whose performance interferes with the effectiveness of an organization should be eliminated from that organization. Such a system of reward and punishment is easy to apply if a man's work can be evaluated in terms of the number or quality of units produced. In most situations, however, the objective evaluation of individual performance is either not possible or extremely difficult. In such cases evaluation usually becomes a matter of a supervisor's over-all judgment of how well an individual does his job. In the past, this supervisory judgment of worker effectiveness was most commonly made on an informal basis. If the supervisor thought a man should be promoted, demoted, or eliminated, he either took or recommended the appropriate action.

With the growth of industry and the parallel growth of personnel departments, it became apparent to some personnel workers that a more formalized or more systematic procedure for evaluation of worker performance was needed. It seemed logical to assume that if supervisors were required to be more analytical in their evaluation the results should be more valid. Thus, procedures were developed for obtaining measures of the extent to which the worker possessed each of a number of traits which were considered by management to be related to success on the job.

In this connection, one of the more interesting research products of World War II was the forced-choice performance rating method developed in the Personnel Research Section of the Adjutant General's Office. Officer ratings obtained on forced-choice rating forms were reported to be substantially less biased and more valid than comparable data obtained by other rating methods (25, 80, 99). Subsequent to the war, the development of forced-choice rating forms for industrial supervisors has been favorably reported by Richardson (66), and Seeley (77) reports success in the use of the method in constructing rating forms for Naval Air Ground School instructors.

While the general procedures for the development of forced-choice rating forms in the reported experiments are similar to those previously employed, certain basic differences in methodology are apparent. When the Human Resources Research Center of the Air Training Command was requested to develop a new rating form for Air Force technical school

instructors, the opportunity to attempt to get answers to certain methodological questions was presented. This bulletin reports the findings of the resulting investigation. Moreover, since forced-choice rating methods are of interest only if they can be shown to be superior to other methods, in the following two sections of this report research findings on forced-choice are compared with those reported for other rating methods.

## CONVENTIONAL RATING PROCEDURES

### General

A conventional rating procedure, as the term will be used here, is intended to mean any rating procedure which requires the rater to indicate where the individual being rated stands on one or more good-poor dimensions. This indication may be made by checking one of several statements which describe various degrees of a given trait or it may be accomplished by making a check mark along a line, one end of which is identified with possession in high degree and the other possession in low degree of the trait or behavior being rated. The simple ranking of employees according to their over-all merit, while widely used, is not included here as a conventional rating procedure.[1]

All conventional rating procedures have one thing in common: it is possible for the rater to tell whether a given check mark on the rating form is going to have a favorable or an unfavorable effect on the ratee's total score.

However, these conventional rating forms have a number of advantages. They are simple, and hence easy to construct. They require little motivation on the part of the rater. They can be quickly filled out and easily scored. Also, claims are made that they provide a convenient medium through which supervisors can deal more effectively with their subordinates. The supervisors are said to be obliged to observe their workers more closely. The competed forms can be used by the supervisor in discussing workers' strong and weak points with them. While these claims seem reasonable, the actual effect of the use of conventional rating procedures on production or morale has yet to be adequately evaluated.

---

[1] Where only a few employees are to be rated, or for experimental use, ranking may provide an economical and relatively satisfactory way of rating personnel. In rating situations involving a number of small groups, however, the average ability and the distribution of abilities within groups may differ sufficiently that the same numerical ranks from different groups represent widely different capabilities.

There is not universal agreement as to the desirability and effectiveness of rating procedures. Harrell (30) says that published descriptions of rating systems are like published descriptions of bridge hands in that they usually report only the instances in which the system worked successfully. Pockrass (57) states, "With the possible exception of individuals who are primarily interested in selling their own pet rating schemes, there is fairly uniform agreement that efforts to measure the employee's service value satisfactorily have been relatively unsuccessful."

A more detailed analysis of the weaknesses of conventional rating procedures may be best presented under the following headings: Validity, Reliability, Error of Leniency, Halo Effect, and Other Aspects. The emphasis in this discussion will be on the use of conventional rating procedures for teacher evaluation.

## Validity

Validity coefficients for rating forms are not always presented in the publications which describe their use. The reason for this dearth of validating statistics is not hard to find. If the situation in which the rating forms are used were such that valid production records or work samples were available as criteria, it probably would not be necessary to use the rating forms. Nearly everyone would agree that, in general, measures of production are to be preferred over subjective ratings. It is in those situations in which production measures are lacking that the principal use has been made of merit-rating devices. Hence, there is seldom a criterion against which the validity of the rating scores can be checked.

Psychologists and others who have been associated with the use of rating forms have found it tempting to assume, in the absence of validating information, that the scores which are yielded by the rating forms have validity. This is an assumption that needs investigation, despite the apparent reasonableness of the idea that a supervisor's opinion of a worker's effectiveness should be highly related to the worker's actual effectiveness.

A number of substitutes for validation against production criteria are described by Driver (16). They include: comparison of scores obtained from rating forms with those obtained from psychological tests, or with those obtained from other rating devices; analysis of the distribution of ratings; and analysis of the presence or absence of halo effect. In addition to these three, supervisor and peer rankings have also been used as criteria. While certain inferences can be drawn from the results of these procedures, they do not permit the prediction of the relationship which exists between rating scores and productive effectiveness on the job.

In discussing validity of rating forms, Cooper (13) states, ". . . it can be said that it is possible to develop rating forms that are statistically reliable, but that one cannot help doubting their validity as devices for measuring occupational proficiency or other performance." Butsch (12), with respect to ratings of general teaching ability, agrees that "Correlation studies have, in general, failed to reveal any significant relationship between [ratings of] general teaching ability and training, scholarship, intelligence, experience, age, salary, credits earned or professional tests." Although this generalization does not have a direct bearing on the relationship between ratings of teaching ability and "actual teaching ability," it tends to throw suspicion on the validity of such ratings. Knudsen and Stephens (41) conclude, from a survey of teacher rating devices which were being used in school situations, "In most instances validity of the device is implied in the assumption that those who furnished the items included were competent to select those traits that make up teaching effectiveness." This assumption is probably just as prevalent in other situations in which ratings are made as it is in teaching situations.

A series of studies conducted at the University of Wisconsin was concerned with the general topic of the measurement of teaching ability. These studies are of special interest here because they deal specifically with the relationship between student growth in subject-matter performance and various judgments of teacher effectiveness. In one of these, Rolfe (68) studied student learning in connection with two three-week units of work in the social studies in one- and two-room rural schools. He found that rating scores as assigned by "experienced and competent supervisors" were correlated .36 to .43 with pupil growth. Rostker (70), in another study dealing with student learning of social studies, found that the relationship between scores on supervisory rating scales and teaching ability as indicated by student progress was insignificant. He concluded that such rating scales should be used only with much discretion. In connection with a factor analysis of the Rolfe and Rostker data, Hellfritzsch (31) concluded that: "teacher rating scales, although frequently used to evaluate the effectiveness of a teacher, are only slightly related to observed pupil growth in social studies. The relationship does not appear to be large enough to warrant using supervisory ratings as a secondary criterion in studies dealing with the measurement of teaching ability, where teaching ability is properly conceived in terms of the ability to promote pupil growth." LaDuke (42), in another of the Wisconsin studies, also found that ratings by superintendents and supervising teachers did not agree with the criterion of student learning.

Baird and Bates (4) studied the ratings by 128 principals of 571 teachers in Detroit. Pupil growth was determined by means of point scores based on the norms of certain standardized tests. A correlation coefficient of only .135 was found between the ratings and this criterion. Taylor (88) found that the correlation between estimations of teaching ability of instructors and progress of pupils in reading and arithmetic was very slight. Pupil progress in arithmetic correlated .018 with the

teacher's estimated ability to teach arithmetic, while progress in reading correlated .241 with estimated ability to teach reading.

In a non-teaching situation Stockford and Bissel (86) found the relatively low correlation of .22 between an objective measure of work performed by mechanics under various supervisors and the ratings of the supervisors by their department heads. In contrast with this, the length of time that the department heads had known each of the rated supervisors correlated .59 and the rated social stimulus value of the personality of the supervisors being rated correlated .65 with the rating scores. This would suggest the possibility that rating scores do a better job of indicating the social acceptability of the one rated than they do of rating his job performance.

From the information which has been presented here, certain generalizations are possible with respect to the validity of performance ratings in general, and of those of instructors in particular. These studies suggest that all ratings need to be considered suspect until the relationship between the ratings and objective measures of job performance has been demonstrated. The data strongly suggest that there is very little relationship between ratings as assigned by public school supervisors and instructor performance as measured in terms of student learning.

## Reliability--Rater Agreement

With respect to rating forms, several kinds of reliability measures are possible. Most commonly the reliability which is given in the published description of a rating form is either a correlation between ratings and reratings of the same individual by the same rater, or the correlation between the scores assigned by different raters using the same form. While it is possible to compute the odd-even reliabilities of most conventional rating devices, this statistic is not usually reported in the technical descriptions of the rating forms.

Knudsen and Stephens (41), in a study of reports of 57 rating devices used in academic situations, found that for 40 of these devices no evidence of reliability was presented.

Butsch (12) presents some reliabilities for same rater-same scale, same rater-different scale, and different rater-same scale. In some of the cases the correlation coefficients are in the .90's, but correlations in the .60's and .70's are more typical. In an industrial situation, Driver (16) found that the correlations of year-to-year ratings by the same raters on 14 traits ranged from .59 to .86. Paterson (56) found month-to-month ratings of the same workers by the same foreman to correlate from .76 to .87. Correlation coefficients between different foremen's ratings of the same workers ranged from .33 to .90.

Cooper (13) reports a correlation of .58 between the ratings of department store salespersons made by two judges. He found ratings of interviewers by their two immediate supervisors to correlate .76. The author concludes that (a) the reliability of the instrument depends not so much upon the form used as upon the situation in which it has been employed; (b) the same kind of rating blank varied in reliability when used in situations involving different kinds of workers; and (c) simple over-all rating proved to be as reliable as an elaborate rating form.

It is difficult to generalize from the published reliabilities since these may not be a random sample of the reliabilities of rating forms in general. The consideration of the evidence on the reliability of rating forms should not cloud the issue with respect to validity. If the ratings lack validity--and the possibility that this is often the case is strongly suggested by the information which is available--the securing of high reliabilities has very little meaning.

## Error of Leniency

Guilford (28) defines "error of leniency" as "a constant tendency that many raters have in common . . . to rate all individuals whom they know above average in certain traits." Kneeland (39) uses the term to describe the tendency of raters to rate well above the midpoint of the scales used. The midpoint in this case is intended to be identified with the average individual. As used here the term "lenient" will have the same meaning as given to it by Kneeland.

Butsch (12) states that most raters rate too high and that this has the effect of producing badly skewed distributions. Richardson (66) points out that the tendency to over-rate apparently increases for every year that a graphic or variant of the graphic system is in operation.

Fry (25) describes the situation which existed prior to the time that the Army's earlier general efficiency rating was replaced by the forced-choice Form 67-1. The former rating placed over 98 per cent of the officers in the excellent categories. Forty-nine per cent of these were in the highest or superior bracket with only a little over 1 per cent in the middle (very satisfactory) group.

Stockford and Bissel (86) found that ratings tended to fluctuate according to whether or not the supervisors were required to discuss the ratings with the individuals rated. When the regular company method of sending the ratings directly to the personnel department without discussions with the individuals rated was used, the mean score on a scale of 100 was 60, with a standard deviation of 21. When experimental ratings were conducted two weeks later, with the rating to be discussed with those rated, the mean rating increased to 84, with a standard deviation of 14. This difference of 24 points between the two means was highly significant (C.R. = 20).

Kneeland (39) showed that the tendency to be lenient exists even when the rater has no apparent reason for being lenient. Fourteen hundred customers were asked to rate sales clerks on five qualities related to job performance (interest in customer, merchandise information, display of merchandise, courtesy, and alertness). The customers rated the salespeople on a 10-point scale on each trait. The average rating on all traits was 6.89. Seventy-five professional shoppers gave a mean rating of 6.01 to the same sales clerks.

Richardson (66) points out that some raters over-rate more than others, with the result that their ratings are not comparable with those made by other supervisors or executives. Stockford and Bissel (86) found that differences in leniency between raters were so great that all of the employees working for the four most severe raters were rated lower than the poorest ratings given by the two most lenient raters. Tiffin (92) accounts for departmental differences in mean ratings in terms of actual discrepancies in merit and differences in standards or interpretation. He suggests that employee ratings should be compared only with those of other employees from the same department rather than with ratings obtained from the plant as a whole. Tiffin's point that the differences in mean ratings of different departments may be partially accounted for by actual differences in the merits of the employees is acceptable on logical grounds. However, in view of the questionable validity of most rating scores the relationship which exists between interdepartmental discrepancies in merit rating scores and in the qualities of job performance must remain indeterminate.

Evans (18) states that the validity of merit ratings is likely to be jeopardized whenever raters react emotionally to something in the rating situation. These emotional reactions on the part of the raters probably have the effect of raising the scores of the individuals rated. Evans believes that the following kinds of feelings on the part of the rater may affect the manner in which he rates: (a) Feelings concerning his inadequacy to make the appraisal. This would include insufficient knowledge of the procedures, of the performance of some or all of the ratees, or inability to rate an employee on some of the rating factors. (b) Feelings of doubt concerning the fairness and accuracy of the rating method. This would include the conviction that the picture will be distorted by some statistical means, by the exclusion of important attributes or the inclusion of unimportant ones. In addition he may lack knowledge of the consequences of his choosing a phrase as "most typical" or "least typical". (c) Feelings of suspicion about what may happen to him (the rater) as a result of the ratings. (d) Feelings of concern for what may happen to his people as a result of the ratings.

Thus there seems to be little doubt, in most rating situations, of the presence of tendencies to rate leniently. These undoubtedly have the effect of lowering the validities of the ratings obtained. The construction of rating devices in such a way as to minimize the leniency phenomenon would seem to be a desirable objective.

## Halo Effect

Thorndike (91) has stated that even a very capable foreman, employer, teacher, or department head is unable to treat an individual as a compound of separate qualities and to assign a magnitude to each of these qualities independently of the others. Rugg (71) has suggested that this inability to rate separate qualities results from tendencies to rate or judge men in terms of a general mental attitude toward them and the domination of this mental attitude toward the personality as a whole over attitudes toward particular qualities.

This phenomenon, which manifests itself in intercorrelations between ratings assigned to supposedly separate traits and in correlations between trait scores and total scores, has come to be known as the "halo effect." Richardson (66) points out that with conventional rating procedures it does not matter which or how many traits or aspects of behavior are listed, since general halo makes it nearly impossible to get a clear picture of a man's strong and weak points. He believes that halo effect results from the failure of conventional rating procedures to separate the reporting of work performance from the evaluation of that performance, and that the evaluation of work performance should therefore be established by statistical means instead of being left to the individual rater's judgment or caprice. He states that the tendencies to over-rate and to evince bias for or against an individual employee are so deep-seated that psychometric techniques must be set up to counteract them.

## Other Aspects of Conventional Rating Procedures

Arbitrary selection of traits. Conventional rating forms usually include provisions for rating on a number of traits or categories of behavior. Actual analyses of the job behavior expected of a man have seldom been made and tested out. Barr (6), from an analysis of 209 teacher rating scales from 46 states, concluded that: (a) a great variety of terms are used to characterize teaching and teaching ability; (b) items are generally highly subjective and ill-defined; (c) content and organization vary widely; (d) social and personal traits surpass, both in frequency and consistency of mention, all other traits enumerated in the study. Knudsen and Stephens (41) found that, in the majority of the 57 teacher rating devices surveyed, the method used for selecting traits to be rated was either individual judgment or the selection of items from other forms.

The use in rating scales of traits or categories of behavior other than those which are really typical or pertinent for a particular job may be one of a number of factors which make for lower validity of rating scores. Flanagan (23) has attacked this problem by use of the "critical incident" technique. This technique includes the collection, from qualified observers, of reports of actual incidents of extremely effective or extremely ineffective job behavior. These incidents are assessed

as to relative frequencies of occurrence and degree of "criticalness." From the resulting data, rating scales that cover only the significant aspects of job behavior can be constructed.

Recall of pertinent behavior. If a supervisor is to do an effective job of rating a subordinate on a certain category of behavior, he must be able to recall all the important employee performances which are related to this behavior category. He must then evaluate each of these performances and arrive at a quantitative summary of them. Obviously this is very difficult, if not impossible, for the average supervisor, or for anyone else. This argues in favor of Flanagan's (23) use of statements of specific behaviors rather than general traits in the construction of rating forms. It also indicates the need for continuous observation and recording by the supervisor of behavioral events which might affect ratings.

Amount of training time required. When rating procedures are found not to work as effectively as is expected, it seems to be common practice for the sponsors of the rating form to explain the failure in terms of lack of training on the part of the raters. Richardson (66) states: "Although training of raters has improved the quality of ratings to some extent, it seems evident that the conventional rating scales demand too much training time. Acceptable ratings have been obtained only as a result of continuous, expensive training of raters. Actually the shoe is on the other foot--a good rating procedure should not only require a minimum of training, but should in itself be a good training device."

## Summary

1. A conventional rating procedure has been defined as one in which the rater can identify the good-poor dimensions. The obviousness of these dimensions makes it possible for the rater to tell what effect a given check mark on the rating form will have on the ratee's total rating score. The rater can raise or lower the ratee's total score as he sees fit.

2. The validity of rating forms i not usually presented in the published descriptions of the forms. One reason is that in those situations in which rating is used there is seldom an adequate standard with which the rating scores can be compared. In the case of instructor rating it has been possible in some cases to compare supervisor ratings with gains in subject-matter knowledge acquired by students. These relationships are characteristically very slight. There is some evidence to indicate that rating scores may be more closely related to the social acceptability of the person rated than it is to his job proficiency.

3. Several kinds of reliability coefficients have been reported in connection with studies of rating forms. A high degree of relationship between the ratings that two different raters give a worker is probably

more important than a high degree of relationship between two ratings by the same rater. Adequate reliabilities are sometimes but not always present with conventional rating procedures. Adequate reliability, however, cannot be accepted as a substitute for adequate validity.

4. A widely prevalent phenomenon in connection with rating has been called the error of leniency. This refers to the tendency to give above-average ratings to the majority of those rated. Such a tendency may impair validity. Over-rating has been observed under a wide variety of conditions but is said to be more exaggerated if the rater is in doubt as to what effect the rating is going to have. Even if it were possible to overcome the propensity of raters to give high ratings, this would not necessarily insure valid ratings.

5. The inability of raters to rate separate qualities independently has been called halo effect. According to some, the presence of this effect argues for additional training of the raters.

6. The execution of periodic ratings may require that the rater be able to remember all the pertinent behaviors which should be considered in passing judgment on an employee's proficiency. He must then be able to evaluate these and arrive at a rating for a trait or an area of behavior. Such remembering and evaluating constitutes a task which the rater may not be capable of carrying out effectively.

7. The criticism has been made that good rating using conventional rating methods requires an impracticable amount of rater training.

## FORCED-CHOICE RATING PROCEDURES

### General

Whether or not the term "forced-choice" is the most appropriate name for the rating procedure which is to be discussed here is debatable. At any rate, it has become almost a household phrase among applied psychologists during the years since World War II. A "forced-choice" rating form is one in which the rater is required to make a series of choices, from groups or blocks of descriptive statements, of those statements which are most (and/or least) descriptive of the person being rated.

While this forced-choice procedure as now used was developed by the Personnel Research Section of the Adjutant General's Office in 1945, it cannot be thought of as without roots in previous psychological research. It was not, for instance, the first method to make use of behavioral statements rather than trait designations. Nor was it the first to use statements which had been systematically evaluated prior to inclusion. Richardson and Kuder (67) describe the development in 1933 of a rating form for salesmen. The procedures used and the rationale underlying them bear an obvious parental resemblance to those perfected later in

the AGO while Richardson and Kuder were on the staff of the Personnel Research Section. Since the war Richardson has developed a number of forced-choice rating forms for the rating of industrial supervisors. He believes that forced-choice ratings come closer to meeting the necessary requirements of a sound rating system than do conventional methods. These requirements Richardson (66) lists as: (a) it should be geared to the needs of the individual organization; (b) it must be reliable; (c) the results of a rating must be expressed in numerical terms; (d) the results must be useful for administration as well as for counseling and training; (e) the content must include elements of the job which have been found to be significant; (f) the ratings must be as free from bias and prejudice as possible; (g) some means must be built into the device to counteract the tendency to rate too high; (h) the form must be easy to fill out; (i) the method should involve a check on the care with which the form is filled out; and (j) the system must be practical in the sense that results can be obtained, recorded, evaluated, and summarized economically.

It should be pointed out that the forced-choice method does not meet completely all the requirements which Richardson has specified. In and of itself its usefulness in the counseling and training of the rater's subordinates has not been demonstrated. However, an additional sheet on which the rater may indicate in a systematic fashion the individual's strong and weak points can be appended to the scale. A copy of the information recorded on this extra sheet may be retained and used by the rater in the counseling and training of his workers. Space can also be provided on this sheet for comments in the rater's own words; this gives the rater a chance to vent any feelings which he may have and which he thinks may not be adequately expressed in the rating form proper. The use of such a procedure in connection with a forced-choice instructor evaluation report has been reported by Seeley (76). As to Richardson's statement that the rating method should involve, if possible, ways of checking on the care and skill with which the form has been filled out, there has been no published information on how this can be accomplished.

A word of caution seems necessary here. While the study reported in the following section of this bulletin has as one of its aims a comparison between conventional and forced-choice rating methods, it is difficult to make a fair comparison from the available research literature. The materials on the deficiencies of conventional rating methods presented above grew out of some 30 years of experience with them. Experience with forced-choice rating methods is very limited. The few published research reports on forced-choice rating tend to emphasize those aspects of it that correct the deficiencies of older methods. A review of these reports thus yields an unavoidable emphasis on the superiorities of the forced-choice method. It could well be that, as more is learned about the results of forced-choice ratings under operational rather than experimental conditions, some of these apparent superiorities will vanish. It is also quite conceivable that, as experience with forced-choice ratings accumulates, psychologists may find that the method has its own unique deficiencies.

## Rationale of Forced-Choice

According to Richardson, a basic assumption of the forced-choice method is that merit rating is best broken down into two distinct phases: (a) reporting the job performance of a man, and (b) evaluating the record or estimate of job performance. He contends that conventional rating procedures force the rater into the necessity of mixing evaluation and reporting, to the detriment of the latter.

Another important feature of the rationale of forced-choice rating is that having to choose between two or more statements forces a critical judgment which is not usually called for by conventional rating procedures. Richardson believes that the rater must ignore to some extent his general impression of a man and think back to specific instances of his work behavior.

In constructing a forced-choice rating form, statements are usually obtained from comments that supervisors make about workers. It is considered to be good practice to leave the statements in their original form as nearly as possible. A minimum of editing is believed to make for ease of understanding on the part of the raters who use the scale.

The analyses which have been made in the building of forced-choice rating forms have indicated that comments which supervisors make about their workers are not of equal significance. Of all the favorable and unfavorable comments which are made in written reports, only a few are highly discriminating between effective and ineffective workers. The forced-choice method attempts to pair discriminating and non-discriminating statements so that the rater must decide which is more descriptive of the individual under consideration.

The matching of statements within blocks according to favorable appearance decreases the possibility of over-rating. The usual pressures toward lenient rating cannot operate when the rater is unable to identify which statements contribute to high scores.

## Validity

One of the principal justifications made for forced-choice rating is that it results in better validities than can be obtained with conventional rating procedures (65, 80, 104). It must be noted that these validities, as in the case of the few reported for conventional rating procedures, are <u>not</u> correlations of forced-choice scores with some measure of performance on the job, but rather are comparisons with ranks or scores that have been assigned to workers by the same supervisors, using different methods.

Wherry (104) reports a study conducted by the Personnel Research Section of the Adjutant General's Office in which five different officer

efficiency reporting methods were investigated. One of the five methods emphasized the forced-choice type of item. This method was found to have consistently higher validities than the four which were more conventional in nature.

Richardson (66) reports validities for forced-choice rating forms ranging from .67 to .74. It should be pointed out that Richardson made use of a purified criterion. That is, persons on whom a reliable rating external to the forced-choice rating could not be obtained were removed from the criterion group. Increasing the reliability of the criterion in this manner undoubtedly has the effect of increasing the computed validities of the forced-choice rating forms somewhat, but the validities would probably still be very substantial even if the criterion had not been purified.

## Reliability--Rater Agreement

Richardson (65) reports reliabilities ranging from .69 to .97 depending on which kind of reliability coefficient was computed. Two odd-even reliabilities, one of .93 and one of .96, are reported. A reliability of .97 was obtained from reratings of the same workers by the same raters using the same forced-choice reporting form. When the raters did the rerating on different forced-choice forms, reliabilities of .93 and .97 were obtained. A relationship of .69 was found between the forced-choice ratings of one rater and those of another rater when different forced-choice forms were used by each rater.

Seeley (77) reports a corrected odd-even reliability of .77 for a forced-choice rating form developed for use in evaluating Navy instructors.

## Error of Leniency

Although one of the principal merits of forced-choice rating methods is supposed to be resistance to deliberate effort to give spuriously high ratings, little evidence has been presented on this point. The data reported by Richardson (65) and by Seeley (77) were obtained under experimental conditions in which rater tendencies toward leniency would certainly not be maximum. Sisson (80), however, gives a graphical comparison of data from a conventional rating (Form 67) and from a forced-choice rating (Form 67-1) of Army officers, when these were used to obtain actual rather than experimental ratings. This graphical comparison is presented in Figure 1. While the forced-choice distribution in Figure 1 is somewhat less skewed (because of more cases at low score levels) than that resulting from use of the conventional form, the difference is not impressive.

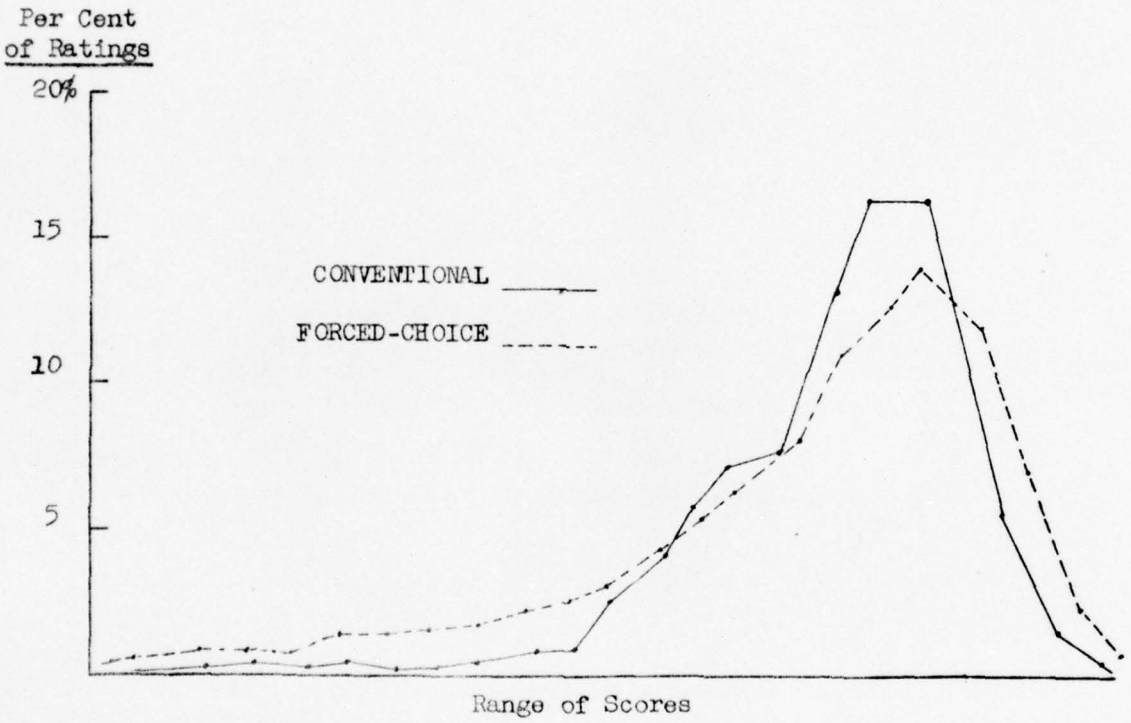


FIG. 1.--Comparative distributions of Army officer scores, with range of scores equated, on two types of rating forms. (After Sisson.)

## Halo Effect

Since forced-choice rating forms do not attempt to evaluate degree of possession of various traits, nor quality of performance in various areas, the concept of "halo effect" may not apply. If, however, the various statements on a form are not so deceptive in their potential effect on total score as the adherents of forced-choice believe, then it would be possible for "halo" to influence the total score.

## Other Aspects of Forced-Choice Rating Procedures

Selection of content. In contrast to the usual content of conventional rating forms (other than those developed by use of the "critical incident" method), the statements included in forced-choice forms are ones that have been made by supervisors about the behavior or characteristics of workers. These statements are then submitted to other supervisors for evaluation prior to inclusion in the form. The only scorable statements that would be included in the form would be those that supervisors agreed on as discriminating between good and poor workers. This should result in considerably more uniform interpretations by raters of the meaning of an item than is possible when abstract traits are being rated.

Rater resistance. There have been unofficial reports that the use of forced-choice methods in the rating of Army officers has met with considerable resistance on the part of the raters. If such resistance turns out, with further experience, to be commonly associated with the use of forced-choice methods, the practical importance of the method is considerably diminished. While the reasons for the reported resistance have apparently not been investigated, the following speculations can be offered.

(a) Raters may resent the fact that they do not know how high or low they have rated a man.

(b) Raters may resent any rating system that really rates. When the rater knows that what he puts on the rating form will affect the career of the person being rated, he commonly tries to avoid ratings that would have an ill effect. The "error of leniency" in conventional ratings is the rater's way out of an unpleasant responsibility. Forced-choice forms, by preventing such deliberate over-rating, may create resentment against the form.

(c) The AGO forced-choice form contained both favorable and unfavorable statements. Raters may resent having to make choices between derogatory statements as descriptive of other officers.

There is some evidence to indicate that regardless of how the raters feel about the ratings this feeling does not invalidate the rating form. Rundquist, Winer, and Falk (74) state that forced-choice items maintained

their relative validities under operational conditions. The total score of the forced-choice Army Efficiency Report had also been shown to maintain its validity when used as the official Army report.

Limited usefulness of results. It is perhaps a legitimate criticism of the forced-choice method that the completed form is of little use to the supervisor in counseling his workers as to their strong and weak points. This may be an inevitable characteristic of any rating method that has as its primary purpose the provision of accurate information regarding the relative abilities of men. Rundquist and Bittner (72) have pointed out: "Ratings which are to serve as a basis for administrative action must yield a valid measure of an individual's performance relative to that of other individuals. The rating system must be specially designed for this purpose; in such a system the rating form or scale takes on particular significance . . . . However, such procedures will probably be found to be of little use in assigning work, in raising morale, or in helping people to improve." As mentioned earlier (page 11), this inadequacy of forced-choice rating procedures can be corrected.

## Summary

1. The forced-choice system of merit rating is one in which the rater is asked to decide which of two or more statements is most descriptive of the individual being rated. The statements from which the rater makes his choices are so arranged that pairs of them have the appearance of being equally favorable things to say about an individual. One of each pair has been found to discriminate between effective and ineffective individuals and the other statement of the pair has been found to be non-discriminating. The blocks of descriptive statements which make up a forced-choice rating form may include either one or two such pairs of statements. If there are two pairs in a block, one pair may possess a different degree of favorableness from the other. In this case the rater is asked to indicate which of the four statements is most descriptive and which is least descriptive of the individual.

2. The forced-choice procedure has been said to separate the reporting of an individual's behavior from the evaluation of that behavior. The rater is not asked to say how much of a certain trait or behavior an individual possesses nor whether it is good or bad to be like that. He has only to indicate which of several statements is more typical. Making this decision may demand more critical judgment than is needed for the completion of a conventional rating form since the rater presumably must think back to actual incidents before making his decision. The language used in a forced-choice rating form will probably be more understandable to the rater since these statements will have been drawn with a minimum of editing from the comments of supervisors themselves.

3. The principal advantages claimed for forced-choice procedure over conventional rating are those of resistance to deliberate efforts

to increase the score and superior validity. Reliability is as high or higher than the reliabilities obtained with conventional rating methods.

4. The reactions of the raters to the execution of forced-choice forms may be a principal factor in determining the eventual position of importance which the forced-choice procedure achieves. If there is excessive resentment on the part of the raters, this may tend to discredit the method. The suggestion has been made that the supervisors be offered opportunity to express their own views in addition to completing the forced-choice form.

## THE DEVELOPMENT OF FORCED-CHOICE FORMS FOR RATING AIR FORCE TECHNICAL SCHOOL INSTRUCTORS

### The Rating Problem

Quality of instruction is crucial in the successful accomplishment of the mission of the Air Training Command. In order to maintain and improve quality of instruction, the Air Training Command is interested in rating methods that will identify, for retention and promotion, those instructors of superior effectiveness, and, for transfer to other duties, those of least effectiveness. Conventional rating methods have proved inadequate for this purpose.

Human Resources Research Center was therefore asked to attempt to develop rating procedures that would better accomplish the desired objectives. Based on the reported experiences of other investigators, the best hope of developing an adequate rating procedure appeared to lie in the use of forced-choice methods.

The previously reported research on forced-choice, however, left a number of methodological questions unanswered. In order that the rating forms and procedures developed for the Training Command be the best possible, these questions needed answering. In addition, it was felt that such methodological information might be of considerable value to other psychologists concerned with rating problems.

### The Methodological Problems

Kind of forced-choice block. The forced-choice blocks reported by Sisson (80), Richardson (65), and Seeley (76) differ in construction. Figure 2 shows a typical block as reported by Sisson.

Two statements have a favorable and two an unfavorable appearance, the rater being instructed to check the most descriptive statement and the least descriptive.

FIGURE 2

Forced-Choice Block--Sisson

| | | MOST | LEAST |
|---|---|---|---|
| A. | Fails to support fellow officers | \|\| | \|\| |
| B. | Oversteps his authority | \|\| | \|\| |
| C. | Gives clear and concise directions | \|\| | \|\| |
| D. | Very exacting in all details | \|\| | \|\| |

FIGURE 3

Forced-Choice Block--Richardson

| MOST | LEAST | |
|---|---|---|
| A | A | Does not play office politics |
| B | B | Is a never-tiring worker |
| C | C | Never reverses a decision |
| D | D | Has difficulty formulating ideas into words |
| E | E | Often absent or tardy |

Richardson's form, illustrated in Figure 3, adds a fifth statement that is neutral in appearance and non-discriminating.

FIGURE 4

Forced-Choice Block--Seeley

a__Able to maintain discipline

b__Has a good memory

c__High degree of efficiency

d__His suggestions and advice are extremely valuable

The form developed by Seeley (see Figure 4) contains only favorable statements, with instructions to the raters to check the two statements that are most descriptive of the person being rated.

Since each of these three forms was developed under different conditions, it is not possible, from the published data, to make valid comparisons among them. Nor is it known that other ways of combining statements into blocks might not be superior to those above.

This study was designed in such manner as to provide comparisons among the above forms. In addition, three new forms were developed and included in the experiment.

The preference index. Sisson (80) refers to the preference index or value of a statement variously as (a) "the extent to which people in general tend to use [it] in describing other people," (b) "general favorableness," and (c) as "the tendency of raters to mark people high or low on the particular behavior item." The formula used for obtaining the index can be related only to this last definition, i.e., $\frac{\Sigma fw}{\Sigma n} \times 100$, where w is the weight on the 5-point scale in terms of which each item was rated.

Richardson (65) lists as separate characteristics of a statement (a) "generally judged favorableness or unfavorableness of the stated behavior" and (b) "popularity, (preference-value, or more explicitly use-frequency) of the element." He does not specify the operations for obtaining indices of either.

Seeley's (76) definition is "preference index, i.e., an average rating or measure of popularity for each phrase as used to describe instructors." He computed this index for individual items by taking half the sum of the mean scores of "best" and "poorest" instructors. These mean scores were obtained from ratings, on a 5-point scale, of the degree to which the items described the instructors. This is essentially the same procedure used by Sisson.

Semantically, it would appear that "extent to which people tend to use a statement" is a different thing from "apparent favorableness" and that neither of these would be measured by the operations described. Therefore, in the present experiment, an attempt was made to get ratings of the "favorableness" of each statement and to compare these with the "preference index."

## Outline of Project Plan

The following steps constituted the general plan of the project:

(1) Collection of a large number of statements describing the performance of instructors.

(2) Collection of supervisor performance rankings of instructors for a sample of the Air Force technical school instructor population.

(3) Obtaining ratings of applicability of the descriptive statements to certain instructor personnel selected on the basis of the supervisor rankings.

(4) Obtaining ratings of favorableness for the descriptive statements.

(5) Computation of preference indices, favorableness indices, and discrimination indices on the basis of the data obtained in steps (2), (3), and (4).

(6) Construction of a number of different kinds of forced-choice rating forms.

(7) Trying out the experimental rating forms on one portion of the population of Air Force technical school instructors, developing scoring keys, and cross-validating on another portion of the same population.

## Execution of the Project Plan

1. Collection of statements describing instructor performance. The basic material from which forced-choice rating forms are constructed is composed of a large number of statements which relate to the performance of the particular job for which the rating form is being designed. Therefore, the first concern of this project was the collection of statements that were, or might be, descriptive of the performance of Air Force technical school instructors.

One of the principal sources of such statements was written remarks of instructor supervisors about the performance of various instructors. These had been written into the space provided for "comments" on a previous instructor rating form. In addition to this source, statements were taken from other instructor rating forms and from rating forms used for rating personnel other than instructors.

A survey of the resulting list revealed that there were a great many more statements with favorable than with unfavorable tone. To equalize this difference, some of the favorable statements were reworded so as to be unfavorable. For example, if a statement indicated that the instructor had behaved in some desirable way, it could easily be reversed by saying that the instructor did not behave in this way. The list, with these changes, totaled 949 statements.

2. Collection of performance rankings. In order to establish a criterion for use in evaluating the statements which had been collected, the immediate supervisors of the technical training instructors at Chanute Air Force Base, Illinois, were asked to rank their instructors as to over-all performance. Only those supervisors were included who had under their supervision at least 5 and not more than 20 instructors.

The supervisors were asked to pick from a list of their instructors the individual they considered to be the most competent of the group and to indicate his rank as "1." They were then asked to give the lowest rank to the individual they considered to be the least competent of the group. Following this, they were instructed to identify the second most competent instructor, the second least competent instructor, etc., until ranks had been assigned to all the instructors in their group whom they had known for a period of at least two months. Assurance was given that the rankings were to be used solely for experimental purposes.

In order to provide some check upon the reliability of these supervisor rankings, it was decided to collect similar rankings from the instructors themselves. That is, from the list of all instructors in his group each instructor was asked to cross out his own name and the names of any instructors whom he had known less then two months. The directions for ranking the remaining names were the same as those given to the supervisors. These rankings were converted to standard scores and these were averaged for each man rated.

The ranks assigned independently by instructors and by supervisors were found to correlate .807 ($N$ = 635 instructors), and thus these provide fairly reliable criterion data for the study. This suggests the question "Why not just use supervisor rankings instead of going to the trouble of developing a rating scale?" The answer, of course, is that for small groups of instructors such rankings may misrepresent the relative abilities of instructors in different groups. It is conceivable that all instructors in one group might be superior to all those in another. Rankings would conceal this, while scores from a good rating scale should reflect it.

3. Obtaining ratings of applicability of the statements. On the basis of the rankings assigned by both the instructor supervisors and the instructors themselves, two extreme groups of instructors were picked. One group consisted of those instructors who were given a top ranking by their supervisor and who were also rated above average by their fellow instructors. The other group consisted of those instructors who were given a bottom ranking by the supervisor and who were also rated below average by their fellow instructors. The 949 descriptive statements were divided among four forms. Fifty-four supervisors were asked to use one of these forms in describing each of two specific instructors. The instructors to be described were those who had been identified both by the supervisor and by fellow instructors as most and least effective in their group--but the supervisor was not told this--he was merely asked to indicate on a 5-point scale (from 0 to 4) the degree of applicability of each statement to these particular instructors. This gave 54 ratings of the applicability of each statement, half on the more and half on the less effective instructors.

4. Obtaining ratings of favorableness on the statements. Forty-six instructor supervisors not used in step 3 were asked to make ratings of

the favorableness of each of the 949 statements by indicating on a 5-point scale (0-4, 0 denoting very unfavorable) how favorable each statement was when used with reference to an instructor.

5. Computing preference, favorableness, and discrimination indices. Preference indices were computed for each of the statements by obtaining a measure of the mean descriptiveness or applicability of each statement. That is, the ratings of the extent to which a statement applied to the effective instructors (see paragraph 3 above) were combined with the ratings of the extent to which it applied to ineffective instructors. The mean of these ratings is an indication of the extent to which the entire group was ranked high or low on a particular statement. The range of these preference indices was from .20 (low) to 3.39 (high).

The index of favorableness was the mean favorableness rating for each statement (paragraph 4 above). The range of these indices was from .20 to 3.59.

The index of discrimination was computed by use of the formula $(\bar{D}_H - \bar{D}_L) \frac{pq}{y}$, in which $\bar{D}_H$ is the mean descriptiveness[2] of the statement for effective instructors, $\bar{D}_L$ is the mean descriptiveness[2] of the statement for ineffective instructors, p is the proportion of the total number of instructors in the high group, q is the proportion of the total number of instructors in the low group, and y is the ordinate of the normal curve corresponding to the values of p and q. The range of the resulting discrimination indices was from -1.46 to +1.59. Figures 5 and 6 show the distribution of these discrimination indices for each level of preference and for each level of favorableness.

6. Comparison of preference and favorableness indices. Because of the bimodal distributions of both preference and favorableness indices, the distribution of favorableness indices was divided at the midpoint of the scale (2.00), and the upper and lower halves were correlated with the preference indices of the same statements. The resulting coefficients were -.03 and +.06, respectively. Apparently, as noted below, these are not indices of the same thing.

The bimodality in terms of the preference index apparently resulted from the fact that favorable items tended to be rated relatively high on applicability to good instructors, but not necessarily to be rated extremely low on applicability to poor instructors. Conversely, items with low favorability indices tended to be rated as having low applicability to good instructors, but not necessarily as having high applicability to poor instructors. This tendency among raters produced bimodality in the distribution of preference indices and induced an apparently high relationship between preference and favorability indices ($r = .89$) when data

(Text continues on page 25)

[2]That is, the mean applicability rating as defined in paragraph 3 above.

FIGURE 5

DISTRIBUTION OF DISCRIMINATION INDICES
AT VARIOUS PREFERENCE LEVELS

(DISCRIMINATION INDEX)

| (PREFERENCE INDEX) | -1.60 -1.41 | -1.40 -1.21 | -1.20 -1.01 | -1.00 -.81 | -.80 -.61 | -.60 -.41 | -.40 -.21 | -.20 -.01 | 0 +.19 | .20 .39 | .40 .59 | .60 .79 | .80 .99 | 1.00 1.19 | 1.20 1.39 | 1.40 1.59 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 379 360 | | | | | | | | | | | | | | | | | | |
| 359 340 | | | | | | | | | | | | | | | | | | |
| 339 320 | | | | | | | | | | | | | | | | | | |
| 319 300 | | | | | | | | | | | | | | | | | | |
| 299 280 | | | | | | | | | | 1 | 2 | | | | | | | |
| 279 260 | | | | | | | | 1 | 3 | 9 | 8 | 14 | 4 | | | | | 3 |
| 259 240 | 1 | | | | | | | 2 | 3 | 10 | 29 | 41 | 15 | 9 | 3 | | | 11 |
| 239 220 | | | | | | | | 2 | 5 | 12 | 38 | 65 | 42 | 9 | | 1 | | 17 |
| 219 200 | | | | | | 1 | | | 1 | 7 | 14 | 41 | 23 | 7 | 2 | | | 9 |
| 199 180 | | 1 | | | 1 | 1 | | | | 5 | 2 | 10 | 10 | 5 | 3 | 1 | | 3 |
| 179 160 | | 1 | | 3 | 2 | 4 | | 1 | 2 | | 1 | | 3 | 2 | 3 | | | 2 |
| 159 140 | | 1 | | 2 | 4 | 2 | 3 | 1 | 1 | | 1 | | | | | | | 1 |
| 139 120 | | 1 | 5 | 4 | 7 | 9 | 3 | 3 | 2 | 2 | | | | | | | | 8 |
| 119 100 | | | 1 | 12 | 35 | 24 | 14 | 3 | | | | | | | | | | 9 |
| 99 80 | | | 3 | 27 | 44 | 41 | 28 | 4 | 1 | | 1 | | | | | | | 14 |
| 79 60 | | | | 11 | 34 | 38 | 37 | 5 | 1 | 1 | | | | | | | | 12 |
| 59 40 | | | | | 4 | 7 | 16 | 6 | 4 | | | | | | | | | 8 |
| 39 20 | 1 | | | | | | 1 | 3 | 2 | | | | | | | | | |
| 19 0 | | | | | | | | | | | | | | | | | | |
| | 2 | 4 | 9 | 59 | 131 | 127 | 102 | 31 | 26 | 47 | 96 | 171 | 98 | 38 | 11 | 2 | | n= 94 |

FIGURE 6

DISTRIBUTION OF DISCRIMINATION INDICES
AT VARIOUS FAVORABLENESS LEVELS

(DISCRIMINATION INDEX)

| (FAVORABLENESS INDEX) | -1.60 -1.41 | -1.40 -1.21 | -1.20 -1.01 | -1.00 -.81 | -.80 -.61 | -.60 -.41 | -.40 -.21 | -.20 -.01 | 0 +.19 | .20 .39 | .40 .59 | .60 .79 | .80 .99 | 1.00 1.19 | 1.20 1.39 | 1.40 1.59 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 379 360 | | | | | | | | | | | | | 2 | | | | | 2 |
| 359 340 | | | | | | | | | | | | | 3 | | | | | 3 |
| 339 320 | | | | | | | | | | | | 7 | 6 | 5 | | 1 | | 19 |
| 319 300 | | | | | | | | | | | 3 | 17 | 20 | 8 | 5 | 1 | | 54 |
| 299 280 | | | | | | | | 1 | | 2 | 16 | 44 | 37 | 6 | 1 | | | 107 |
| 279 260 | | | | | | | | 1 | 3 | 14 | 39 | 61 | 18 | 7 | 4 | | | 147 |
| 259 240 | | | | | | | | 1 | 4 | 16 | 28 | 32 | 9 | 4 | 1 | | | 95 |
| 239 220 | | | | | | | | 1 | 3 | 10 | 10 | 5 | 1 | 3 | | | | 33 |
| 219 200 | | | | | | 1 | | 1 | 3 | 3 | | 2 | 2 | | | | | 12 |
| 199 180 | | 1 | | | | | 3 | 1 | 1 | 1 | | 3 | | | | | | 10 |
| 179 160 | | | | 2 | 2 | 7 | 1 | | 1 | | | | | | | | | 13 |
| 159 140 | 1 | 1 | 3 | 9 | 10 | 10 | 3 | 2 | 1 | | | | | | | | | 40 |
| 139 120 | | | 1 | 8 | 17 | 17 | 13 | 3 | 2 | | | | | | | | | 61 |
| 119 100 | | 2 | 2 | 14 | 38 | 36 | 21 | 9 | 2 | 1 | | | | | | | | 125 |
| 99 80 | 1 | | 2 | 21 | 45 | 29 | 38 | 3 | 2 | | | | | | | | | 141 |
| 79 60 | | | | 4 | 13 | 22 | 17 | 6 | 3 | | | | | | | | | 65 |
| 59 40 | | | | | 6 | 5 | 4 | 2 | 1 | | | | | | | | | 18 |
| 39 20 | | | 1 | 1 | | | 2 | | | | | | | | | | | 4 |
| 19 0 | | | | | | | | | | | | | | | | | | |
| | 2 | 4 | 9 | 59 | 131 | 127 | 102 | 31 | 26 | 47 | 96 | 171 | 98 | 33 | 11 | 2 | | n= 949 |

on both favorable and unfavorable items were pooled. However, a negligible relationship is evident between the two indices when data from favorable and unfavorable items are considered separately.

The reasons for this lack of agreement appear to lie in the nature of the two indices. The preference index is the mean degree of applicability of a statement to the entire population (or to the high and low extremes thereof). The same mean degree of applicability could result from statements that differed considerably in their degrees of applicability to the high and low groups. For instance, three statements with applicability (or descriptiveness) mean scores of 4, 3, and 2 for the high group would yield the same preference index when the mean scores for the low group were 0, 1, and 2, respectively. But the purpose of computing such an index is so that statements that appear equally favorable can be paired in the forced-choice blocks. And statements with descriptiveness means of 4 and 0 for the high and low groups, respectively, inevitably appear to be more favorable than statements for which the descriptiveness is the same for both groups. The preference index, being an average, obscures these differences.

The favorableness index, on the other hand, was a direct attempt to ascertain how favorable a statement looked to the supervisors who were ultimately to use the forced-choice form. Since the preference and favorableness indices were dissimilar, and since the latter seemed more likely to represent a statement's appearance of favorableness, the favorableness index was used for the balance of the study.

7. Construction of the forced-choice rating forms. The forced-choice method calls for the pairing, within blocks, of statements that are of equivalent favorableness, but that differ in discrimination. As can be seen from Figure 6, while many such statements were available, the number that could be used depended on how large a difference in discrimination indices was demanded. If this difference were set too high, then only those statements at the extremes of the distribution of discrimination indices within each favorableness level could be used. If it were set too low, the statements might fail to discriminate when put into the forced-choice blocks. In this study a difference of .60 was chosen solely because it was the largest difference that would make available the number of statements necessary for forms of optimum length. However, this leaves an interesting methodological question unanswered:

What should the discrimination difference between two statements be in order to achieve maximum economy in the use of the pool of statements, maximum validity and reliability, and minimum biasability of the final forced-choice form? One might expect that as larger discrimination differences were used reliability would increase, validity might increase, but resistance to bias might decrease. The discrimination difference that will provide an optimum relationship between these desirable characteristics of a forced-choice form needs identification.

Utilizing this discrimination difference of .60, six forced-choice forms were constructed as shown below. The favorableness indices (FI) and discrimination indices (DI) given for each statement are merely illustrative--they are not the correct indices for the statements shown.

FORM A

Seventy-three blocks, two statements per block. There were roughly equal numbers of favorable and unfavorable blocks.

Summary of Directions: Pick the statement which is more descriptive (favorable blocks), or less descriptive (unfavorable blocks).

Sample Blocks:
a. Aim of lesson is clearly presented. (FI 2.78, DI .63)
b. Refrains from spending too much time boasting of his experiences. (FI 2.61, DI .02)

a. May "bawl out" or ridicule a student in the presence of others. (FI .90, DI -.95)
b. Doesn't get to know each student's problems. (FI .87, DI -.34)

FORM B

Thirty-four blocks, three statements per block. One of the three statements had a discrimination index .60 higher than the other two. There were roughly equal numbers of favorable and unfavorable blocks.

Summary of Directions: Pick the statement which is most descriptive and the one which is least descriptive in each block.

Sample Blocks:
a. Does not answer all questions to the satisfaction of the students. (FI 1.43, DI -.20)
b. Does not use proper voice volume. (FI 1.47, DI -.80)
c. Supporting details are not relevant. (FI 1.40, DI -.15)

a. Conducts class in orderly manner. (FI 2.22, DI 1.20)
b. Repeats questions to the whole class before answering them. (FI 2.29, DI .57)
c. At ease before class. (FI 2.35, DI .53)

## FORM C

Thirty-one blocks, four statements per block. All statements had high favorableness indices.

Summary of Directions: Pick the two statements which are most descriptive.

Sample Block:
a. Patient with slow learners. (FI 2.82, DI 1.15)
b. Lectures with confidence. (FI 2.75, DI .54)
c. Keeps interest and attention of class. (FI 2.89, DI 1.59)
d. Acquaints classes with objective for each lesson in advance. (FI 2.85, DI 1.19)

## FORM D

This form was identical with Form C except for the directions.

Summary of Directions: Pick the statement which is most descriptive and the one which is least descriptive in each block.

Sample Block: Same as Form C.

## FORM E

Thirty-two blocks, four statements per block. Two had high and two had low favorableness indices.

Summary of Directions: Pick the statement which is most descriptive and the one which is least descriptive in each block.

Sample Blocks:
a. Fine personal bearing. (FI 3.01, DI 1.21)
b. Adapts himself readily to new duties. (FI 2.98, DI .59)
c. Is not well qualified to instruct in all phases of his subject. (FI .65, DI -.75)
d. Does not put class at ease. (FI .78, DI -.13)

## FORM F

Thirty-six blocks, five statements per block. Two had high and two had low favorableness indices; the fifth statement had a favorableness index midway between the high and low pairs, and a low discrimination index.

Summary of Directions: Pick the statement which is most descriptive and the one which is least descriptive in each block.

Sample Block:
a. Works hard. (FI 3.26, DI 1.39)
b. Somewhat antagonistic about what he is instructed to do. (FI 1.22, DI -.96)
c. Could improve cleanliness of classroom area. (FI 1.89, DI .05)
d. Not willing to adapt to changing situations. (FI 1.26, DI -.30)
e. Can take criticism. (FI 3.30, DI .78)

Form C uses Seeley's (76) method of constructing blocks, Form E uses the AGO method reported by Sisson (80), and Form F follows Richardson (64). Forms A, B, and D are rather obvious alternative constructions developed for this experiment.

8. Experimental testing of the six rating forms. The six rating forms were experimentally administered at Chanute, Scott, Keesler, Warren, Lowry, and Sheppard Air Force Bases. For purposes of cross-validation, the instructors at the first three bases named were used as one sample and those from the last three bases as another.

In addition to the administration of the forms at the bases indicated, supervisors were asked to rank their subordinates as to their over-all effectiveness as instructors, using the method described on page 20. These ranks were converted into a normalized ranking score. This was necessary so that ranks from groups of different sizes could be given comparable meaning. The procedure, however, makes the assumption that the mean level of performance of all groups is the same. Since many of the groups consisted of but 5 or 6 instructors, such an assumption is almost certainly unwarranted. To the extent that the assumption is unwarranted, these normalized scores will give an inaccurate report of the relative abilities of instructors from different groups. Since these are the criterion scores against which the six rating forms were validated, it can be assumed that the validity coefficients obtained are conservative estimates.

At Scott and Keesler Air Force Bases, supervisors were also asked to fill out the rating forms according to the following directions: "Fill out this rating form as if you were rating your best friend and wanted to make certain that he obtained as high a score as possible." The data gathered in this manner will be referred to hereafter as the "bias experiment."

The number of cases which were obtained by form and base is presented in Table 1.

TABLE 1

NUMBER OF EXPERIMENTAL CASES ACCORDING TO RATING FORM AND BASE

| Base | Form | | | | | | No. |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | F | Raters |
| Chanute | 67 | 77 | 63 | 72 | 69 | 73 | 90 |
| Scott | 46 | 48 | 43 | 49 | 46 | 47 | 56 |
| Keesler | 50 | 56 | 33 | 52 | 53 | 45 | 37 |
| Warren | 42 | 38 | 43 | 44 | 39 | 42 | 26 |
| Lowry | 49 | 51 | 50 | 45 | 47 | 43 | 24 |
| Sheppard | 38 | 46 | 38 | 42 | 47 | 44 | 28 |
| Bias (Scott) | 65 | 66 | 57 | 65 | 65 | 67 | |
| Bias (Keesler) | 25 | 28 | 31 | 28 | 28 | 28 | |

## Discussion and Results

Analysis Procedures. For purposes of analysis, the data collected from Scott, Chanute, and Keesler Air Force Bases were combined and will be referred to as Group I. The combined data from Lowry, Warren, and Sheppard Air Force Bases are designated Group II.

For each group the following procedures were carried out: Using the normalized ranking score as a criterion, the highest one-third and the lowest one-third of the completed forced-choice forms were selected. Graphic item counts were run for each response position[3] for the high and low groups for each of the six kinds of rating forms. These counts were transferred to summary sheets and validity indices were determined by the use of Davis (15) tables.

Five experimental scoring keys were made for each of the six forced-choice rating forms:

Key 1. This key was based on the original indices of discrimination of the statements, in terms of the Chanut Sample (see p. 22). Weights of +1, 0, or -1 were assigned depending on the relative size and sign of the indices.

Key 2. This key was based on item analysis of Group I data. Using Davis item validity indices, response positions having indices above 6 were weighted +1, those below -6 were weighted -1, and those between -6 and +6 were given zero weight. No blocks were scored unless one or more

[3]The term "response position" refers to the possible responses to the statements in the forced-choice form. For example, if a forced-choice block contains three statements and the instructions are to check the most and least descriptive statements, there would be two possible response positions for each statement (most and least descriptive) and six possible response positions in the block.

statements had indices above +12 and one or more statements had indices below -12. No minus weights were used for Form A.

Key 3. This key used item-analysis results, but unit weights were assigned in accordance with the logical relations of the various response positions. Blocks were selected for scoring in the same manner as for Key 2. Opposite responses for the same statement took opposite weights, unless both alternatives had Davis indices between +6 and -6. In the latter case, both response positions were weighted zero. For Form A, Keys 2 and 3 were identical. Since in this form there were only two alternatives per block, the weighting of one alternative positively required assignment of a negative weight to the other.

Key 4. This key was developed in the same manner as Key 2, but used Group II data.

Key 5. This key was developed in the same manner as Key 3, using Group II data. For Form A, Keys 4 and 5 were identical.

In the construction of the four keys based on item-analysis data, it was found that some blocks on each of the six forms did not yield scores that differentiated significantly between the high and low groups. This would seem to indicate that the discriminative power of a statement may be different when considered alone than when considered in comparison with certain other statements. Or it may be that the discrimination index, as determined here, gives only a very rough indication of the relative discriminating power of the statements.

Table 2 shows the extent of shrinkage of each form when non-discriminating blocks are eliminated after analysis of Group I data.

TABLE 2

EFFECT OF ITEM ANALYSIS ON LENGTH OF FORMS[a]

| Form | Original Length | | Scorable Length | | Per Cent Shrinkage |
|---|---|---|---|---|---|
| | Blocks | Statements | Blocks | Statements | |
| A | 72 | 144 | 32 | 64 | 55 |
| B | 34 | 102 | 16 | 48 | 53 |
| C | 31 | 124 | 26 | 104 | 16 |
| D | 31 | 124 | 20 | 80 | 35 |
| E | 32 | 128 | 30 | 120 | 6 |
| F | 36 | 180 | 34 | 170 | 6 |

From these data it would appear that, if two- or three-choice blocks are to be used, considerable shrinkage should be anticipated. Whether it is necessary to correct for this by starting with longer experimental forms

[a]Based on Group I data and Key 2.

should be determinable from the relationships of the number of statements in a form to the form's coefficients of validity and reliability. These relationships are explored below.

Validities. As mentioned earlier, the criterion in this study consists of rankings by instructor supervisors of their subordinates' overall competence as instructors. It is recognized that such a criterion is not necessarily a valid measure of the actual effectiveness of instructors. A more valid measure would probably be based on the relative effectiveness of instructors in producing changes in the behavior and attitudes of students. Such an ultimate criterion is particularly difficult to obtain in a training situation such as that in Air Force technical schools, since the many different courses present an extremely wide range of difficulty, and the teaching of each instructor is commonly limited to a small phase of one course.

The principal reason for the present project was the need for a rating form or forms that would accurately report, for administrative uses, what supervisors believed to be the relative effectiveness of their instructors. In terms of this limited objective the use of rankings as the criterion is considered justified.

Table 3 presents the correlation coefficients which were obtained between the criterion and scores from each of the five keys on each of the six forms for the two groups of data.

An examination of these data reveals a rather unusual situation. It would be expected that the keys developed on Group I would produce higher validities when they were used on Group I than they would produce when they were used on Group II. This expectation is not borne out by the data. Of the 17 correlations obtained using Group I keys (Keys 1, 2, 3) on Group II data, all but one of the comparable correlations are higher for Group II than they are for Group I. The situation is decisively reversed when Group II keys (Keys 4, 5) are used on Group I data. In this case, all of the correlations are lower for Group I than for Group II, i.e., are lower on cross-validation.

This presence of consistently higher validities for Group II than for Group I casts doubt on the Group I criterion data. An examination of the criterion for both groups reveals that Group I contains a predominance of small groups. That is, many of the groups which the supervisors in Group I ranked contained four, five, or six instructors. With the Group II data, the average size of the groups ranked was considerably larger. As discussed on page 28, when rankings are converted to normalized standard scores, the resulting scores may misrepresent the relative abilities of instructors who come from groups that differ in mean ability. The smaller the groups, the more probable it is that such misrepresentation will occur. It is considered, therefore, that cross-validation data obtained when Keys 1, 2, and 3 are used on Group II may be a better indication of the relative validities of the six experimental

TABLE 3

VALIDITIES OF FORCED-CHOICE FORMS USING VARIOUS KEYS[a]
(Correlations with rankings)

Group I

| Form | KEY 1* | 2* | 3* | 4 | 5 |
|---|---|---|---|---|---|
| A | .486 | .609 | -- | .501 | -- |
| B | .551 | .609 | .595 | .553 | .502 |
| C | .447 | .573 | .562 | .451 | .464 |
| D | .523 | .619 | .602 | .565 | .545 |
| E | .548 | .567 | .573 | .535 | .548 |
| F | .492 | .547 | .537 | .450 | .494 |

Group II

| Form | KEY 1 | 2 | 3 | 4* | 5* |
|---|---|---|---|---|---|
| A | .500 | .636 | -- | .670 | -- |
| B | .475 | .577 | .525 | .681 | .624 |
| C | .655 | .703 | .704 | .714 | .674 |
| D | .643 | .680 | .663 | .712 | .708 |
| E | .584 | .537 | .549 | .620 | .616 |
| F | .583 | .564 | .594 | .664 | .609 |

[a] Keys 2 and 3 were derived on Group I data, Keys 4 and 5 on Group II data. Except for being based on different samples, Keys 2 and 4 are comparable, as are Keys 3 and 5. The asterisk in certain column headings denotes that the correlations listed were based on the same sample used to derive the key in question and are therefore to some degree spurious.

forms than the cross-validation data resulting from Keys 4 and 5 on Group I.

It should be noted, however, that Keys 1, 2, and 3 are derived exclusively from Group I and, therefore, may be expected to suffer from the limitations of this group. It may be assumed that the deficiencies existing in Group I are likely to affect all six experimental forms in the same way and hence will not alter the relative order of the validities which would result if most of the deficiencies did not exist. It seems advisable, then, to consider all the cross-validation data in evaluating the various experimental forms. These data are shown in Table 4. Inspection of Table 3 yields little evidence that the relative validity of the various keys differed markedly over the six forms. Therefore, averaging of the data by keys cannot be considered to obscure "key by form" interactions.

As validity is only one of the important characteristics of a forced-choice rating plan and as the validity under operational conditions remains to be evaluated, this is not to be considered a final ranking of the relative value of the forms. It is interesting to note that the two forms containing all favorable statements in each block, Forms C and D, show consistently higher cross-validation coefficients than any other forms.

The validities of all forms exceed a correlation of .404 obtained between a sample of 442 scores from the graphic rating form previously used to rate instructors and supervisor rankings. It may be noted also that the magnitudes of the average validity coefficients do not appear to be primarily determined by the lengths of the various forms.

It is planned that Keys 2, 3, 4, and 5 will be evaluated later in this study, after data are collected under operating conditions. Key 1 will probably be dropped since it is considered inferior to the others, having been developed prior to the item analysis and having been derived from the discrimination indices for the statements taken singly (not in blocks) and based on only part of the Group I criterion data (those from Chanute AFB). For parts of the balance of this study it was necessary to choose one key. Any one of the four keys (2, 3, 4, or 5) might properly have been chosen. Preference for a completely empirical key narrowed the choice to either Key 2 or Key 4. Key 2 was available earlier in the study and some analysis had already been completed with it. Hence, when only one key is used in the balance of this bulletin it is Key 2 rather than Key 4.

The Bias Experiment. The purpose of this experiment was to determine the relative resistance of the various forms to deliberate attempts to give high scores. The participants were instructed to consider that they were rating their best friend and to check the form so that he would get as high a score as possible.

TABLE 4

RESULTS OF CROSS-VALIDATION OF SIX FORCED-CHOICE FORMS

| Form | No. of Blocks | No. Blocks Scorable on Key 2 | No. Blocks Scorable on Key 4 | Statements Per Block | Kinds of Statements Per Block | Instructions | Average[a] Cross-Validity Grp II[b] | Average[a] Cross-Validity Grps I & II[c] |
|---|---|---|---|---|---|---|---|---|
| C | 31 | 26 | 19 | 4 | All favorable | Check 2 most descriptive statements. | .686 | .607 |
| D | 31 | 20 | 24 | 4 | All favorable | Check the most and least descriptive statements. | .662 | .622 |
| A | 73 | 32 | 43 | 2 | Both statements favorable or both unfavorable | Check most descriptive in favorable blocks, least descriptive in unfavorable. | .614 | .578 |
| F | 36 | 34 | 36 | 5 | Two favorable, two unfavorable, and one neutral statement | Check the most and least descriptive statements. | .580 | .539 |
| E | 32 | 30 | 31 | 4 | Two favorable and two unfavorable | Check the most and least descriptive statements. | .558 | .551 |
| B | 34 | 16 | 23 | 3 | All favorable or all unfavorable | Check the most and least descriptive statements. | .528 | .527 |

[a]Computed by use of Fisher's z transformation.

[b]Using Keys 1, 2, 3 on Group II.

[c]Using Keys 1, 2, 3 on Group II and Keys 4 and 5 on Group I.

The distribution statistics for the six forms under both experimental and bias conditions are presented in Table 5.

The upward shift of the mean from experimental to bias conditions can be taken as one measure of the extent to which raters have succeeded in biasing the scores. In order to compare the magnitudes of these shifts, the differences between the means for Group II on Key 2 and the biased means on Key 2 were divided by the standard deviations of the Group II distributions. This gave the following ranking of forms in order of decreasing resistance to bias:

| FORM | $\frac{M_{Bias} - M_{Grp\ II}}{\sigma_{Grp\ II}}$ |
|---|---|
| C | .44 |
| B | .60 |
| A | .71 |
| D | .74 |
| E | .74 |
| F | .74 |

A clearer picture of the over-all effects of biasing instructions on the distributions of scores from the various forms can be gained from Figures 7 through 12. Forms E and F are clearly less bias-resistant than the other forms. The bias distribution for Form E is quite similar to the distribution reported by Sisson for the same kind of form under conditions of actual use (Figure 1). Forms E and F are the only forms of the six in which favorable and unfavorable statements are included in the same block. This probably increased the biasability of the form because the rater is almost certain that "most" answers to favorable statements and "least" answers to unfavorable statements are likely to increase the total score. Forms with unmixed blocks do not offer such information to the rater.

Form C, which was shown to have the smallest shift of mean score under instructions to bias and was one of the two highest in validity, also yields the most normal appearing distribution of scores under these conditions.

It seems probable that the bias obtained in this experiment is maximum and that the actual amount of effort to bias that would be exerted when a form was in regular use would be somewhat less. The distributions that could therefore be expected when these forms were put into use should be somewhere between those obtained here under regular experimental conditions and those obtained under instructions to bias.

Reliability. Reliability coefficients for the six forms are presented in Table 6. The procedures used in their computation differ somewhat from the conventional odd-even method.

(Text continues on page 41)

TABLE 5

DISTRIBUTION STATISTICS FOR SIX FORCED-CHOICE FORMS UNDER EXPERIMENTAL CONDITIONS AND UNDER DIRECTIONS TO ATTEMPT TO GIVE AS HIGH A SCORE AS POSSIBLE (BIAS)

| Form | Key[a] | Group I | | | Group II | | | Bias | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | N | Mean | SD | N | Mean | SD | N | Mean | SD |
| A | 1 | 163 | 36.8 | 9.4 | 127 | 36.4 | 10.8 | 90 | 42.5 | 6.6 |
| A | 2 | 163 | 16.7 | 6.2 | 127 | 17.1 | 6.2 | 90 | 21.5 | 4.0 |
| A | 4 | 163 | 23.2 | 8.1 | 127 | 22.9 | 9.4 | 90 | 28.4 | 5.1 |
| B | 1 | 181 | 34.3 | 7.8 | 133 | 34.1 | 9.1 | 94 | 39.6 | 5.2 |
| B | 2 | 181 | 2.4 | 10.7 | 133 | 1.2 | 11.1 | 94 | 7.9 | 7.0 |
| B | 3 | 181 | 17.7 | 5.8 | 133 | 17.3 | 6.2 | 94 | 21.0 | 3.8 |
| B | 4 | 181 | 2.4 | 13.2 | 133 | 2.1 | 15.8 | 94 | 11.5 | 9.9 |
| B | 5 | 181 | 25.1 | 7.2 | 133 | 24.9 | 8.6 | 94 | 29.8 | 5.3 |
| C | 1 | 138 | 32.5 | 8.2 | 129 | 33.3 | 8.5 | 88 | 36.2 | 5.7 |
| C | 2 | 138 | -1.5 | 13.1 | 129 | -1.5 | 13.2 | 88 | 4.3 | 8.5 |
| C | 3 | 138 | 18.4 | 6.5 | 129 | 18.1 | 6.7 | 88 | 21.4 | 4.1 |
| C | 4 | 138 | 0.2 | 10.3 | 129 | 0.3 | 12.7 | 88 | 5.9 | 6.9 |
| C | 5 | 138 | 17.9 | 5.4 | 129 | 17.5 | 7.2 | 88 | 20.9 | 3.7 |
| D | 1 | 173 | 2.4 | 10.6 | 129 | 2.2 | 10.0 | 93 | 9.3 | 6.3 |
| D | 2 | 173 | 0.8 | 14.0 | 129 | -0.2 | 13.6 | 93 | 9.8 | 9.7 |
| D | 3 | 173 | 21.5 | 7.7 | 129 | 21.1 | 7.6 | 93 | 27.1 | 5.2 |
| D | 4 | 173 | 2.1 | 15.0 | 129 | 0.9 | 17.4 | 93 | 13.3 | 11.3 |
| D | 5 | 173 | 27.7 | 8.1 | 129 | 27.2 | 9.0 | 93 | 34.3 | 6.2 |
| E | 1 | 168 | 21.6 | 17.1 | 132 | 18.8 | 23.2 | 93 | 38.4 | 13.4 |
| E | 2 | 168 | 10.9 | 25.2 | 132 | 8.5 | 26.7 | 93 | 28.3 | 18.0 |
| E | 3 | 168 | 20.0 | 20.3 | 132 | 16.9 | 22.1 | 93 | 30.3 | 13.2 |
| E | 4 | 168 | 13.9 | 24.5 | 132 | 12.4 | 28.6 | 93 | 31.1 | 16.7 |
| E | 5 | 168 | 15.6 | 20.1 | 132 | 18.7 | 23.6 | 93 | 33.0 | 13.7 |
| F | 1 | 165 | 20.4 | 22.0 | 128 | 18.7 | 24.9 | 95 | 36.2 | 11.5 |
| F | 2 | 165 | 16.9 | 29.1 | 128 | 15.8 | 30.0 | 95 | 37.9 | 13.6 |
| F | 3 | 165 | 16.0 | 20.2 | 128 | 15.3 | 21.7 | 95 | 30.3 | 9.2 |
| F | 4 | 165 | 19.2 | 29.4 | 128 | 15.8 | 35.0 | 95 | 42.5 | 15.7 |
| F | 5 | 165 | 20.5 | 23.0 | 128 | 19.3 | 25.6 | 95 | 36.6 | 11.4 |

[a]For Form A, Keys 2 and 3 are identical, and Keys 4 and 5 are identical.

FIG. 7

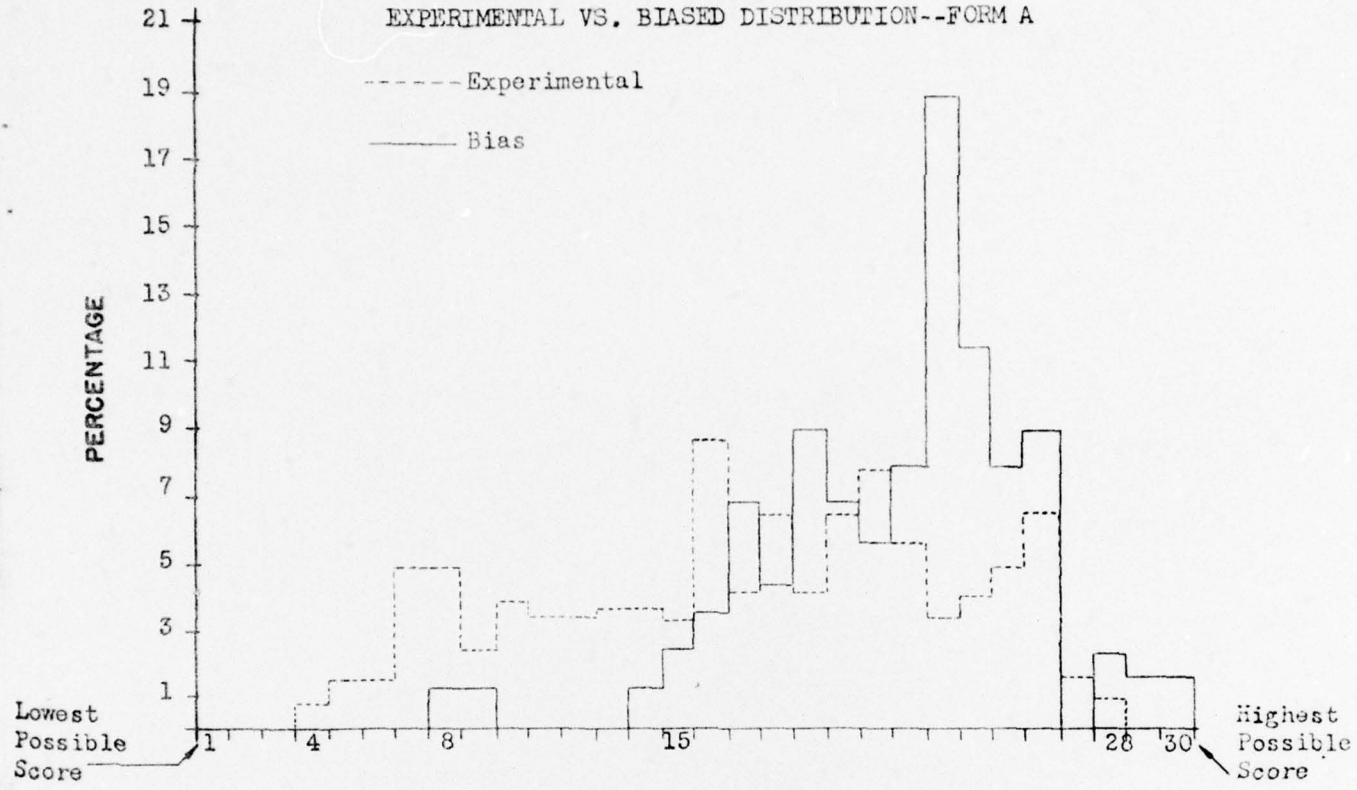
EXPERIMENTAL VS. BIASED DISTRIBUTION--FORM A
Experimental
Bias
PERCENTAGE
21
19
17
15
13
11
9
7
5
3
1
Lowest Possible Score
1
4
8
15
28
30
Highest Possible Score

FIG. 8

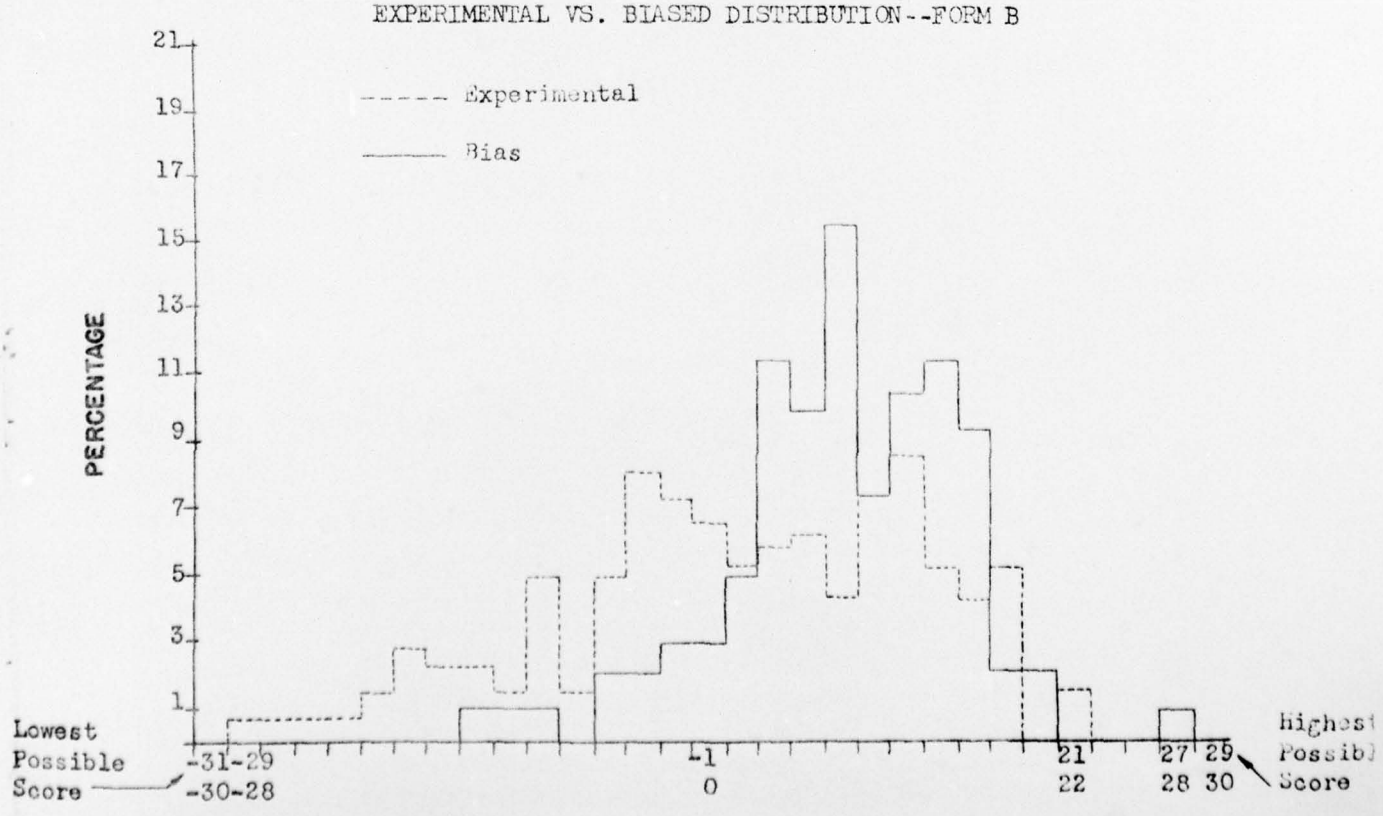
EXPERIMENTAL VS. BIASED DISTRIBUTION--FORM B
Experimental
Bias
PERCENTAGE
21
19
17
15
13
11
9
7
5
3
1
Lowest Possible Score
-31-29
-30-28
-1
0
21
22
27 29
28 30
Highest Possible Score

FIG. 9

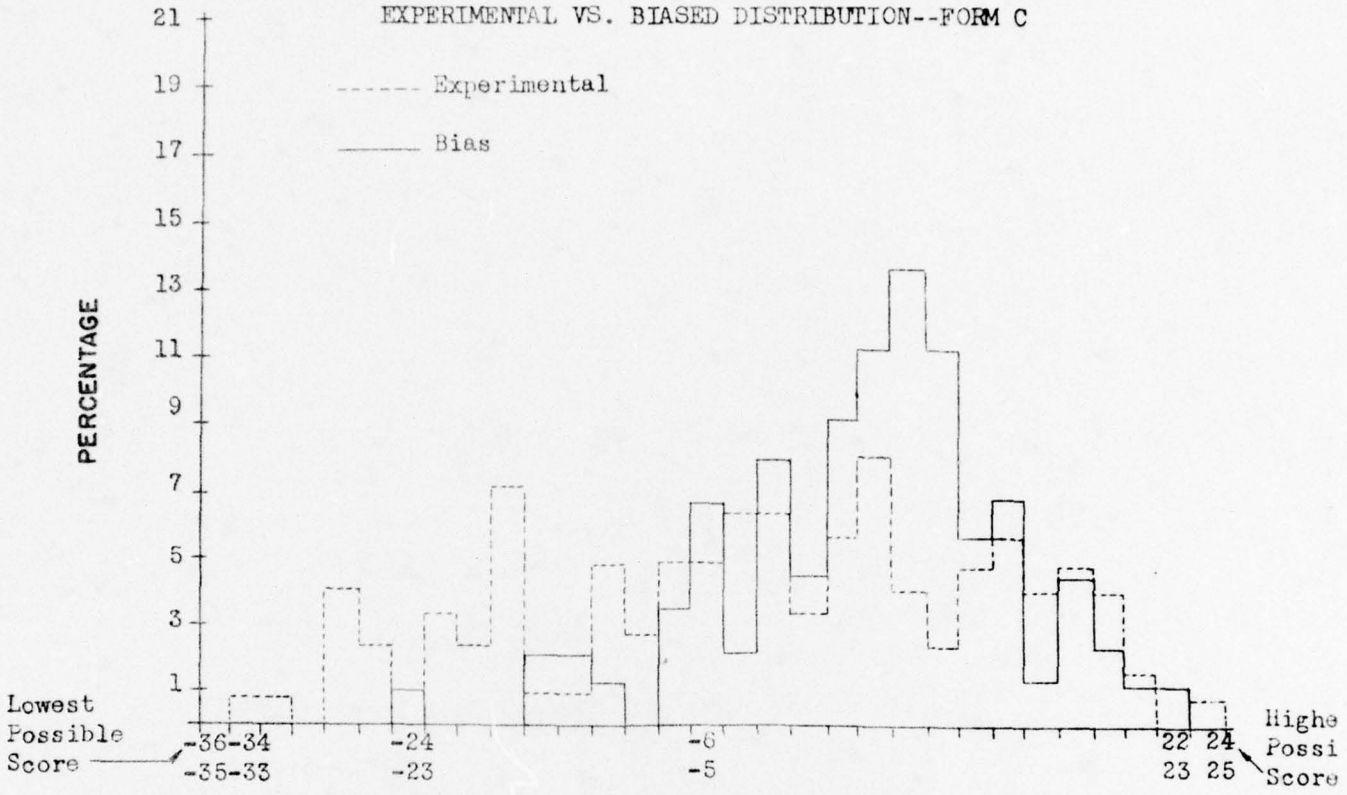
EXPERIMENTAL VS. BIASED DISTRIBUTION--FORM C
Experimental
Bias
PERCENTAGE
21
19
17
15
13
11
9
7
5
3
1
Lowest Possible Score
-36-34
-35-33
-24
-23
-6
-5
22 24
23 25
Highe
Possi
Score

FIG. 10

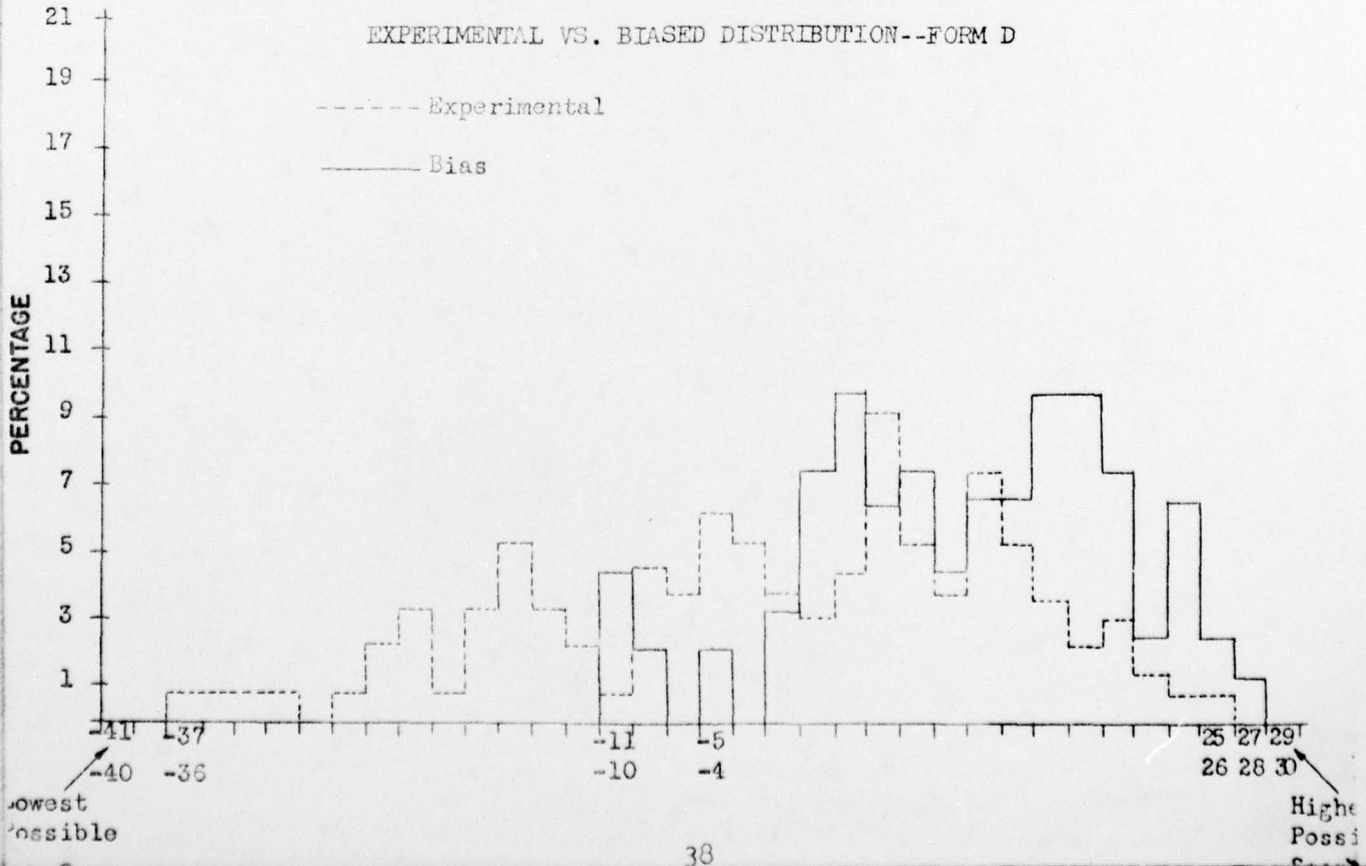
EXPERIMENTAL VS. BIASED DISTRIBUTION--FORM D
Experimental
Bias
PERCENTAGE
21
19
17
15
13
11
9
7
5
3
1
-41
-40
-37
-36
-11
-10
-5
-4
25 27 29
26 28 30
owest
ossible
e
Highe
Possi

FIG. 11

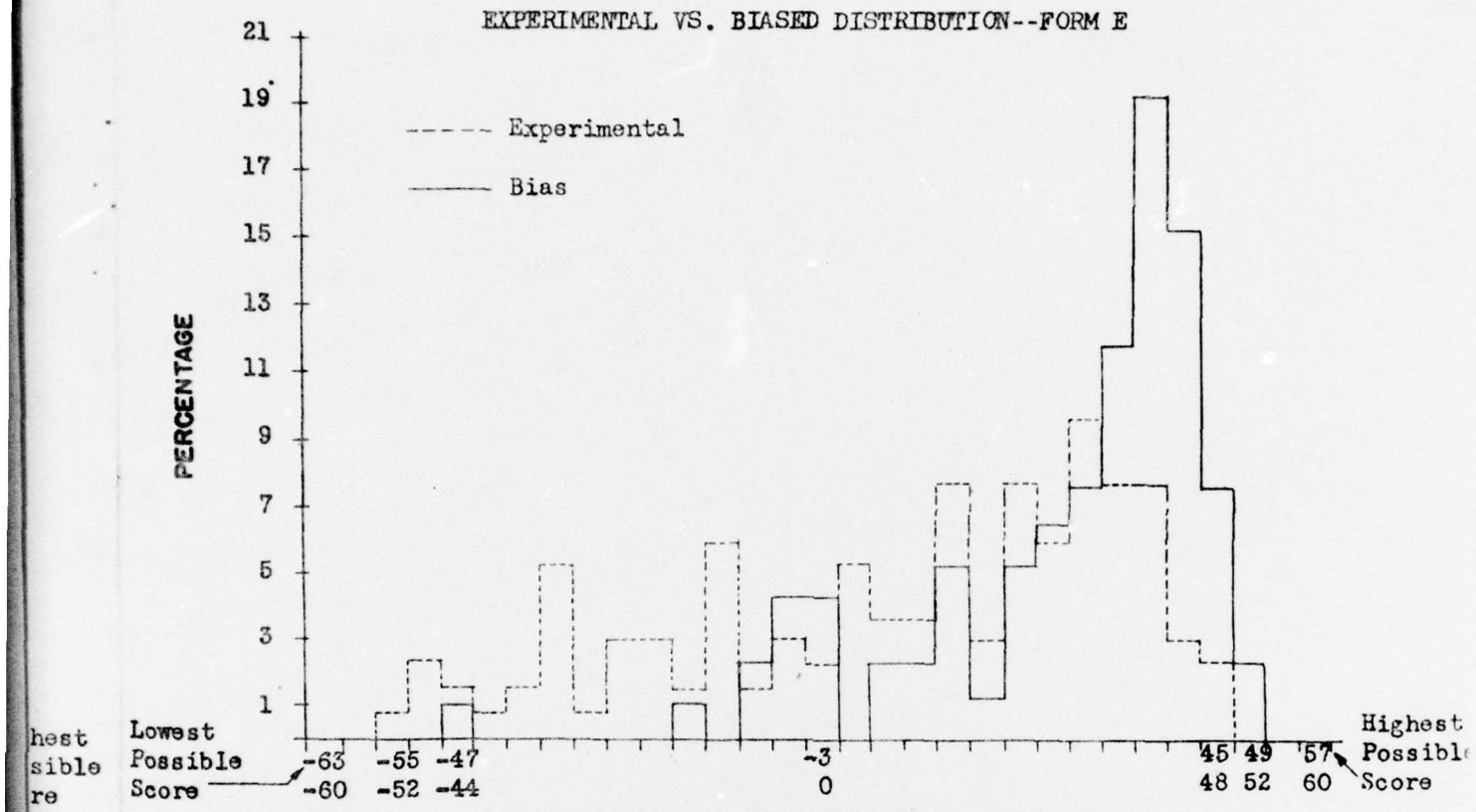
EXPERIMENTAL VS. BIASED DISTRIBUTION--FORM E
Experimental
Bias
PERCENTAGE
21
19
17
15
13
11
9
7
5
3
1
Lowest Possible Score
-63 -55 -47 -3 45 49 57
-60 -52 -44 0 48 52 60
Highest Possible Score

FIG. 12

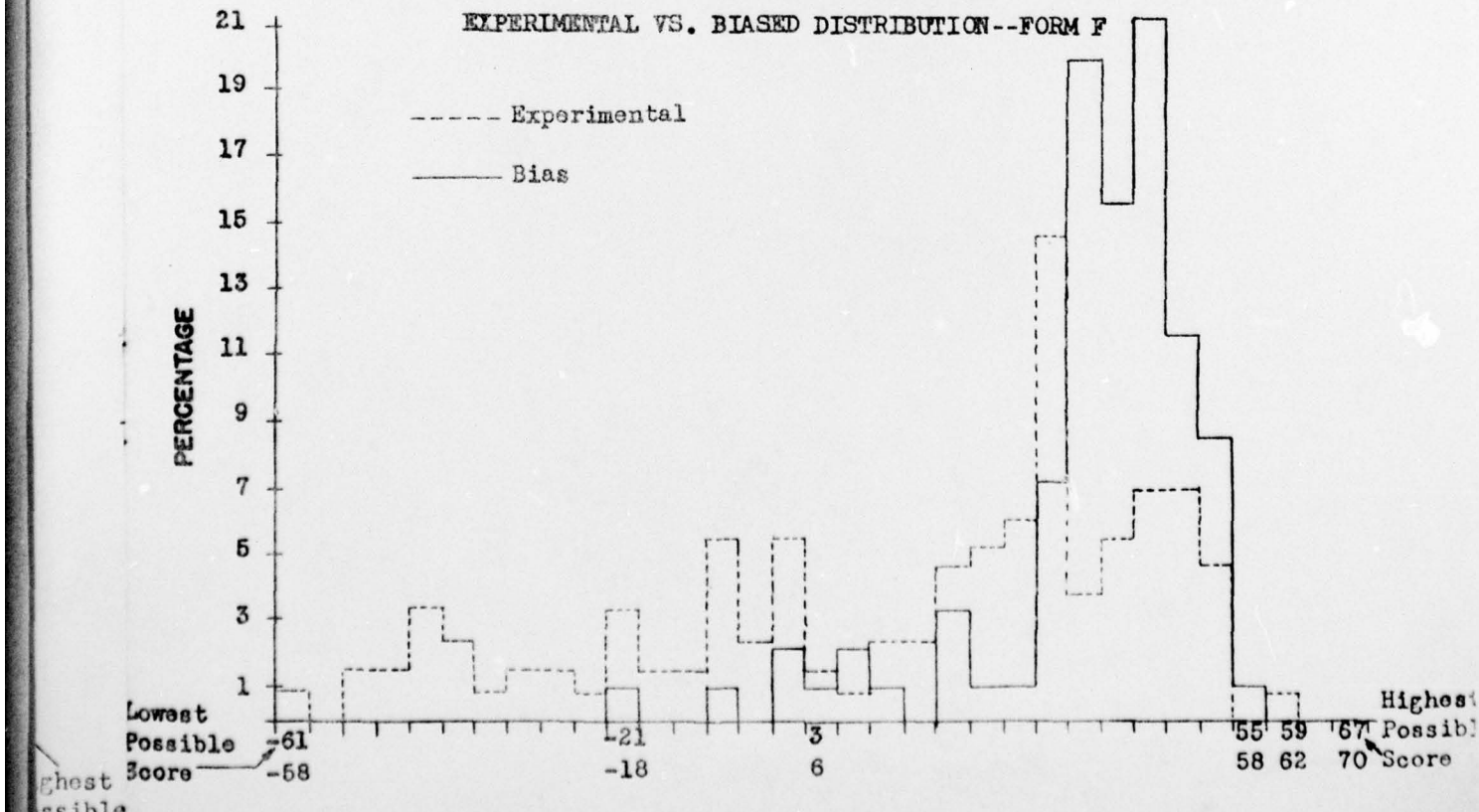
EXPERIMENTAL VS. BIASED DISTRIBUTION--FORM F
Experimental
Bias
PERCENTAGE
21
19
17
15
13
11
9
7
5
3
1
Lowest Possible Score
-61 -21 3 55 59 67
-58 -18 6 58 62 70
Highest Possible Score

TABLE 6

RELIABILITY COEFFICIENTS FOR INSTRUCTOR DESCRIPTION FORMS

| 1. Form | 2. Number of Statements | 3. Group I[a] | 4. Group II[a] | 5. Bias | 6. Group I[b] | 7. Group II[b] |
|---|---|---|---|---|---|---|
| A | 64 | .785 | .817 | .635 | .907 | .922 |
| B | 48 | .657 | .737 | .451 | .872 | .908 |
| C | 104 | .888 | .897 | .730 | .926 | .934 |
| D | 80 | .834 | .853 | .674 | .914 | .925 |
| E | 120 | .941 | .952 | .932 | .958 | .966 |
| F | 170 | .923 | .959 | .833 | .923 | .959 |

[a]Split-half coefficients stepped up by Spearman-Brown formula.

[b]Coefficients resulting when all forms are equated in length to the longest form (170 statements) by Spearman-Brown formula; e.g., for Form A, which has 64 statements, the reliability for Group II would be .922 if there were 170 such statements.

a. The following data were used in computing reliabilities:

(1) Group I and Group II answer sheets were combined and scored with Key 2.

(2) The Bias Group was scored with Key 2.

b. The results of the earlier item analysis of the six instructor description forms were used in splitting the blocks of each form into two comparable groups as follows:

(1) A validity index was obtained for each block in a given form.

(2) The blocks were arranged in rank order with respect to this validity index.

(3) An odd-even split of the blocks was then made from this rank order.

(4) Minor adjustments in the split were made to balance on the criterion of number of choices scored per block.

c. Coefficients of correlation were obtained between the odd and even blocks for each form.

It is evident from Table 6 that the three forms with less than 100 statements each (A,B,D) have the lower coefficients (Cols, 3,4). When a correction is made for the length of the forms, all coefficients reach satisfactory levels (Cols. 6,7). It would therefore seem to be advisable, when constructing forced-choice forms made up of 2- or 3-choice blocks, to include in the experimental forms enough extra statements to compensate for the excessive shrinkage that takes place with forms so constructed.

When forced-choice forms are constructed in the manners reported here, using discrimination differences around .60, it appears that final forms about 100-120 statements in length should yield reliability coefficients around .90.

The smaller coefficients shown for the Bias data (Col. 5) can be accounted for by the shrinkage of the variance of the distributions of scores under instructions to bias. When corrected for the difference in variance these coefficients are not significantly different from those obtained for Group II (Col. 4).

More meaningful kinds of reliability than that presented above would be the agreement between different raters using the same form, different raters using different forms, and the same rater using different forms. The first two of these are not obtainable in the technical instructor situation, since in the great majority of groups there is but one immediate

supervisor. Data on the inter-form reliability using the same raters are being collected under operational conditions and will be reported later.

Supervisors' Ratings of the Relative Desirability of the Various Forms. A number of supervisors who participated in the forced-choice experiment were asked to state which of the six forms (or however many forms they used in rating their group of instructors) they liked best, next best, etc. If a supervisor rated six or more instructors, he made use of all six of the rating forms. If he had less than six instructors, he used as many of the forms as he had instructors. In other words, each instructor was rated on only one form, but as many of the forms as possible were used by each supervisor in rating his instructors. This resulted in from 80 to 83 rankings of each form.

The mean ranks in the order of most desirable to least desirable for each of the six forms were:

1. Form C 2.89
2. Form A 2.90
3. Form F 3.12
4. Form E 3.40
5. Form D 3.69
6. Form B 4.66

It should be recognized that these data reflect only the ranked desirability of the forced-choice forms presented. Such a ranking cannot show intensity of feeling, i.e., it is conceivable that two forms having adjacent ranks might be widely separated on a like-dislike continuum. It is also possible that the raters may have very much liked or very much disliked all forms.

The most useful conclusion from the data would seem to be that if forced-choice forms are to be used, then forms arranged as are Forms C and A are somewhat less likely to be disliked than are the others.

## SUMMARY AND CONCLUSIONS

1. In this study, 949 statements about various aspects of the performance of Air Force technical school instructors were collected.

2. Preference, favorableness, and discrimination indices of each statement were computed.

3. Six experimental forced-choice instructor rating forms were constructed. These differed from each other in number of statements per block, in homogeneity of blocks with regard to favorableness of statements, or in directions to the rater.

4. The forms were used by instructor supervisors to rate instructors at six Air Force bases. Additional supervisors at two bases completed the forms under instructions to give as high a rating as possible.

5. From the data available, five experimental scoring keys were developed.

6. The instructor supervisors were asked to rank the instructors they supervised according to over-all competence as an instructor. Correlations between these ranks and scores on the rating forms were separately computed for the six bases, six forms, and the five keys. The 56 coefficients obtained ranged from .441 to .714. Under conditions of cross-validation, Forms C and D (4-item blocks, all favorable statements) had the highest average validity.

7. When supervisors were instructed to give as high a rating as possible and when biasability was estimated in terms of the resulting mean shift, Form C proved least biasable, with Forms E and F being markedly less resistant to efforts to bias than the other four forms.

8. Reliability coefficients for each form were computed by a modification of the odd-even method. These ranged from .657 to .959. When corrected for differences in lengths of forms, the coefficients ranged from .908 to .966.

9. Supervisors were asked to rank the six forms as to desirability. Forms C and A were about equally preferred over the others.

10. It is concluded that, of the six forms tested here, Form C yields the best over-all results, since it was one of the two highest in validity, was lowest in biasability, had satisfactory reliability, and was one of the two forms best liked, or least disliked, by the raters.

11. The conclusions of the study are limited by the fact that the data were collected in an experimental situation. Forms A, B, C, and D are now in use for the regular rating of Air Force technical school instructors. The resulting data are to be compared with those presented here.

BIBLIOGRAPHY


1. ADAMS, H.F. Validity, reliability and objectivity. Psychol. Monogr., 1936, 47, 329-350.

2. ALBERTY, H.B., and THAYER, V.T. Supervision in the secondary school. Boston: Heath, 1931.

3. ALMY, H.C., and SORENSON, H. A teacher-rating scale of determined reliability and validity. Educ. Adm. Supervis., 1930, 16, 179-186.

4. BAIRD, J., and BATES, G. The basis of teacher rating. Educ. Adm. Supervis., 1929, 15, 175-183.

5. BARR, A.S. An introduction to the scientific study of classroom supervision. New York: Appleton, 1931.

6. BARR, A.S. Measurement of teaching ability. Rev. educ. Res., 1940, 10, 182-184.

7. BOWMAN, E.C. Pupil rating of student teachers. Educ. Adm. Supervis., 1934, 20, 141-146.

8. BRADSHAW, F.F. The American Council on Education rating scale: its reliability, validity and use. Arch. Psychol., 1930, 119, 80 p.

9. BRYAN, R.C. Eighty-six teachers try evaluating student reactions to themselves. Educ. Adm. Supervis., 1941, 27, 513-526.

10. BRYAN, R.C. Why student reactions to teachers should be evaluated. Educ. Adm. Supervis., 1941, 27, 590-603.

11. Bureau of Public Personnel Administration. What's wrong with service (efficiency) ratings? Public Personnel Studies, 1929, 7, 18-28.

12. BUTSCH, R.L. Teacher rating. Rev. educ. Res., 1931, 1, 99-107, 156-157.

13. COOPER, J.H. Rating forms. In Stead, W.H., Shartle, C.L., and others, Occupational counseling techniques. New York: American Book, 1940.

14. CORREVANT, H.E. The purpose and methods involved in teacher rating. Phi Delta Kappan, 1939, 21, 25-29.

15. DAVIS, F.B. Item analysis data, their computation interpretation and use in test construction. Cambridge, Mass.: Graduate School of Education, Harvard University, 1946.

16. DRIVER, R.S. A case history of merit rating. Personnel, 1940, 16, 137-162.

17. DRIVER, R.S. The validity and reliability of ratings. Personnel, 1941, 17, 185-191.

18. EVANS, J.W. Emotional bias in merit rating. Personnel J., 1950, 28, 290-291.

19. EWART, E., SEASHORE, S.E., and TIFFIN, J. A factor analysis of an industrial merit rating scale. J. appl. Psychol., 1941, 25, 481-486.

20. FERGUSON, L.W. The value of acquaintance ratings in criterion research. Personnel Psychol., 1949, 2, 93-102.

21. FLANAGAN, J.C. (Ed.) The aviation psychology program in the Army Air Forces, AAF Aviation Psychology Program Research Report No. 1. Washington, D.C.: U.S. Govt. Printing Office, 1948.

22. FLANAGAN, J.C. An analysis of the results from the first annual edition of the National Teacher Examinations. J. Educ., 1941, 9, 237-250.

23. FLANAGAN, J.C. Critical requirements: a new approach to employee evaluation. Personnel Psychol., 1949, 2, 419-425.

24. FREYD, M. The graphic rating scale. J. educ. Psychol., 1923, 14, 83-102.

25. FRY, J.C. All superior officers. Infantry J., 1948, 63, 21-26.

26. GARRETT, H.E., and SCHNECK, M.R. Psychological tests, methods and results. New York: Harper, 1933.

27. GARVEY, J.A. Rating in theory and practice at the Dennison Manufacturing Company. Amer. Mgmt. Ass., Annual Convention Series, 1926, 43.

28. GUILFORD, J.P. Psychometric methods. New York: McGraw Hill, 1936.

29. HALSEY, G.D. Making and using industrial service ratings. New York: Harper, 1944.

30. HARRELL, T.W. Industrial psychology. New York: Rinehart, 1949.

31. HELLFRITZSCH, A.G. A factor analysis of teacher abilities. J. exp. Educ., 1945, 14, 166-199.

32. HILL, R.L. Efficiency ratings. Personnel J., 1936-37, 15, 330-332.

33. HOLLINGWORTH, H.L. Judging human character. New York: Appleton, 1922.

34. HOPKINS, J.T. Some fallacies and virtues of merit rating. Amer. Mgmt. Ass., Production Series, 124, 30-39.

35. JONES, R.D. The prediction of teaching efficiency from objective measures. J. exp. Educ., 15, 85-100.

36. KELLEY, T.L. Principles underlying the classification of men. J appl. Psychol., 1913, 3, 50-67.

37. KING, J.E. Multiple item approach to merit rating. Amer. Psychologist, 1949, 4, 278 (Abstract).

38. KINGSBURY, F.A. Making rating-scales work. J. Personnel Res., 1925, 4, 1-6.

39. KNEELAND, NATALIE. That lenient tendency in rating. Personnel J., 1928-1929, 7, 356-366.

40. KNOWLES, A.S. Merit rating in industry. Northeastern Univ. Bull., 1940, 1.

41. KNUDSEN, C.W., and STEPHENS, STELLA. An analysis of fifty-seven devices for rating teaching. Peabody J. Educ., 1931, 9, 15-24.

42. LaDUKE, C.V. The measurement of teaching ability. J. exp. Educ., 1945, 14, 75-100.

43. LANDIS, C. The justification of judgments. J. Personnel Res., 1925, 4, 7-19.

44. LEVINE, M. Some concepts of efficiency rating. Personnel Admin., 1942, 5, 14-18; 1943, 5, 8-14.

45. LINS, L.J. The prediction of teaching efficiency. J. exp. Educ., 1946, 15, 2-60.

46. LYNCH, J.M. The psychology of the rating scale. Educ. Adm. Supervis., 1944, 30, 497-501.

47. LYNCH, J.M. Teacher rating trends psychologically examined. Amer. Sch. Bd. J., 1942, 104, 27-28.

48. MAHLER, W.R. Let's get more scientific in rating employees. Personnel, 1947, 23, 310-320.

49. MARKEY, S.C. Consistency of descriptive personality phases in the forced-choice technique. Amer. Psychologist, 1947, 2, 210-211 (Abstract).

50. MARSH, SARAH E., and PERRIN, F.A.C. An experimental study of the rating scale technique. J. abnorm. soc. Psychol., 1925, 19, 383-399.

51. McGINNIS, W.C. Supervisory visits and teacher rating devices. J. educ. Res., 1934, 28, 44-47.

52. MOORE, H. Are employee rating plans the bunk? Amer. Business, 1942, 12, 18, 26-27.

53. MYERS, G.J., JR. Personnel rating. General Electric Rev., 1942, 45, 351-356.

54. National Industrial Conference Board. Employee rating, studies in personnel policy. New York: 1942, 39.

55. ORDWAY, S.H., JR., and LOFFAN, J.C. Approaches to the measurement and reward of effective work of individual government employees. National Municipal Rev. Supplement, 1935, 24, 557-601.

56. PATERSON, D.G. The Scott Company graphic rating scale. J. Personnel Res., 1922-23, 1, 361, 376.

57. POCKRASS, J.H. Common fallacies in employee ratings. Personnel J., 1939-40, 18, 262-267.

58. POSEY, C.W. A new answer to an old problem--shall we rate teachers? Amer. Sch. Bd. J., 1944, 108, 34-35.

59. PROTZMAN, MERLE I. Student rating of college teaching. Sch. & Soc., 1929, 29, 513-515.

60. RADOM, M. Picking better foremen. Factory Mgmt. & Maintenance, 1950.

61. REMMERS, H.H. The college professor as the student sees him. Purdue Univ. Stud. higher Educ., 1929, 11.

62. REMMERS, H.H. To what extent do grades influence student ratings of instructors? J. educ. Res., 1930, 21, 314-317.

63. REMMERS, H.H., and GAGE, N.L. Educational measurement and evaluation. New York: Harper, 1943, 450-485.

64. REMMERS, H.H., and PLICE, M.J. Reliability of ratings at Purdue University. Industr. Psychol., 1926, 1, 717-721.

65. RICHARDSON, M.W. An empirical study of the forced-choice performance report. Paper presented at the 57th Annual Meeting of the American Psychological Association, Denver, Colorado, September, 1949.

66. RICHARDSON, M.W. Forced-choice performance reports: a modern merit-rating method. Personnel, 1949, 26, 205-212.

67. RICHARDSON, M.W., and KUDER, G.F. Making a rating scale that measures. Personnel J., 1933-34, 12, 36-40.

68. ROLFE, J.F. The measurement of teaching ability. J. exp. Educ., 1945, 14, 52-74.

69. ROSTKER, L.E. The measurement and prediction of teaching ability. Sch. & Soc., 1940, 51.

70. ROSTKER, L.E. The measurement of teaching ability. J. exp. Educ., 1945, 14, 6-51.

71. RUGG, H. Is the rating of human character practicable? Educ. Psychol., 1921, 12, 425-438; 1921, 12, 485-501; 1922, 13, 30-42; 1922, 13, 81-93.

72. RUNDQUIST, E.A., and BITTNER, R.H. A merit rating procedure developed by and for the raters. Personnel, 1950, 26, 273-283.

73. RUNDQUIST, E.A., and BITTNER, R.H. Using ratings to validate personnel instruments; a study in method. Personnel Psychol., 1948, 1, 163-183.

74. RUNDQUIST, E.A., WINER, B.J., and FALK, GLORIA H. Follow-up validation of forced-choice items of the army officer efficiency report. Amer. Psychologist, 1950, 5, 359 (Abstract).

75. RYAN, T.A. Merit rating criticized. Personnel J., 1945-46, 24, 6-15.

76. SEELEY, L.C. Construction of three measures for instructor evaluation. Technical Report--SDC 383-1-5, Office of Naval Research, July 20, 1948.

77. SEELEY, L.C. Preliminary validation of the instructors evaluation report. Technical Report--SDC 383-1-9, Office of Naval Research, April 20, 1949.

78. SELLS, S.B., and TRAVERS, R.M.W. Observational methods of research. Rev. educ. Res., 1945, 15, 394-407.

79. SISSON, E.D. The criterion in army personnel research. In Kelly, G.A., New methods in applied psychology, 1947, 17-21.

80. SISSON, E.D. Forced-choice--the new Army rating. Personnel Psychol., 1948, 1, 365-381.

81. STAFF, PRS, Personnel Research and Procedures Branch, Adjutant General's Office. The forced-choice technique and rating scales. Amer. Psychologist, 1946, 1, 267 (Abstract).

82. STARR, R.B., and GREENLY, R.J. Merit rating survey findings. Personnel J., 1939, 17, 378-384.

83. STEVENS, S.N., and WONDERLIC, E.F. An effective revision of the rating technique. Personnel J., 1934, 13, 124-135.

84. STEWART, NAOMI. Methodological investigation of the forced-choice technique, utilizing the officer description and the officer evaluation blanks (OCS classes 461 and 464, Fort Benning, Georgia). Adjutant General's Office, Personnel Research Section, Report No. 701, 6 Jul 45.

85. STEWART, NAOMI. Obtaining officer preference and officer characteristics scale values of adjectives for use in construction of items for the biographical information blank. Adjutant General's Office, Personnel Research Section, Report No. 702, 7 Jul 45.

86. STOCKFORD, L., and BISSEL, H.W. Factors involved in establishing a merit rating scale. Personnel, 1949, 26, 94-116.

87. SYMONDS, P.M. Diagnosing personality and conduct. New York: Century, 1931.

88. TAYLOR, H.R. Teacher influence on class achievement. Genet. Psychol. Monogr., 1930, 7, 81-174.

89. THAYER, V.T. Teacher rating in the secondary schools. Educ. Adm. Supervis., 1926, 12, 361-378.

90. THORNDIKE, E.L. A constant error in psychological ratings. J. appl. Psychol., 1920, 4, 25-29.

91. THORNDIKE, E.L. Fundamental theorems in judging men. J. appl. Psychol., 1918, 2, 67-76.

92. TIFFIN, J. Industrial Psychology. New York: Prentice-Hall, 1947.

93. TROW, W.C. How shall teaching be e[illegible]uated? Educ. Adm. Supervis., 1934, 20, 264-272.

94. TYLER, R.W. Techniques for evaluating behavior. Educ. Res. Bull., Ohio St. Univ., 1934, 13, 1-11.

95. VERNON, P.E. The assessment of psychological qualities by verbal methods. Industr. Res. Bd. Rep., Med. Res. Council, London, 1938, 83, 43-88, 111-121.

96. VON HADEN, H.I. An evaluation of certain types of personal data employed in the prediction of teaching efficiency. J. exp. Educ., 1946-47, 15, 61-84.

97. WADSWORTH, G.W. Practical employee ratings. Personnel J., 1935, 13, 263-269.

98. WALKER, HELEN M. (Ed.) The measurement of teaching efficiency. New York: Macmillan, 1935.

99. WHERRY, R.J. Comparative validity of the WD AGO Form No. 67 and the FCL-2 according to various breakdowns. Adjutant General's Office, Personnel Research Section, Report No. 671, 7 Dec 45.

100. WHERRY, R.J. Comparative validity of the WD AGO Form No. 67 and the FCL-2 according to various breakdowns II; European theater. Adjutant General's Office, Personnel Research Section, Report No. 673, 8 Dec 45.

101. WHERRY, R.J. Construction and scoring of the officer efficiency report OER-A (WD AGO PRT-520). Adjutant General's Office, Personnel Research Section, Report No. 678, 10 Oct 45.

102. WHERRY, R.J. Construction and scoring of the officer efficiency reports. FCL-2a, b, and c. Adjutant General's Office, Personnel Research Section, Report No. 679, 9 Oct 45.

103. WHERRY, R.J. Experimental evidence of the value of ranking as a method of rating. Adjutant General's Office, Personnel Research Section, Report No. 677, 10 Dec 45.

104. WHERRY, R.J. Major study of comparative validity of five periodic officer efficiency reporting methods. Adjutant General's Office Personnel Research Section, Report No. 670, 5 Dec 45.

105. WHERRY, R.J. Major study of comparative validity of five periodic efficiency reporting methods II; European theater. Adjutant General's Office, Personnel Research Section, Report No. 672.

106. WHERRY, R.J., and FRYER, D.H. Buddy ratings: popularity contest or leadership criteria. Personnel Psychol., 1949, 2, 147-159.